

Mixtures of Generalized Hyperbolic Distributions and Mixtures of Skew-t Distributions for Model-Based Clustering with Incomplete Data

Yuhong Wei and Paul D. McNicholas

Dept. of Mathematics & Statistics, McMaster University, Hamilton, Ontario, Canada.

Abstract

Robust clustering from incomplete data is an important topic because, in many practical situations, real data sets are heavy-tailed, asymmetric, and/or have arbitrary patterns of missing observations. Flexible methods and algorithms for model-based clustering are presented via mixture of the generalized hyperbolic distributions and its limiting case, the mixture of multivariate skew-t distributions. An analytically feasible EM algorithm is formulated for parameter estimation and imputation of missing values for mixture models employing missing at random mechanisms. The proposed methodologies are investigated through a simulation study with varying proportions of synthetic missing values and illustrated using a real dataset. Comparisons are made with those obtained from the traditional mixture of generalized hyperbolic distribution counterparts by filling in the missing data using the mean imputation method.

1 Introduction

Finite mixture models presented as powerful and flexible tools for discovering heterogeneity in multivariate datasets. Assuming no prior knowledge of class labels, the application of finite mixture models in this way is known as model-based clustering. As McNicholas (2016a) points out, the association between mixture models and clustering goes back at least as far as Tiedeman (1955), who uses the former as a means of defining the latter. Gaussian mixture models are historically the most popular tool for model-based clustering

and dominated the literature for quite some time (e.g., Celeux and Govaert, 1995; Fraley and Raftery, 1998; McLachlan et al., 2003; Bouveyron et al., 2007; McNicholas and Murphy, 2008, 2010). The multivariate t -distribution, being a heavy-tailed alternative to the multivariate Gaussian distribution, made (robust) mixture modelling based on mixtures of multivariate t -distributions the most natural extension (e.g., Peel and McLachlan, 2000; Andrews and McNicholas, 2011, 2012; Lin et al., 2014). In many practical situations, however, real world datasets exhibit clusters that are not just heavy tailed but also asymmetric; furthermore, clusters can also be asymmetric yet not heavy tailed. Over the few past years, much attention has been paid to non-Gaussian approaches to model-based clustering and classification, including work on multivariate skew- t distributions (e.g., Lin, 2010; Vrbik and McNicholas, 2012; Lee and McLachlan, 2014; Murray et al., 2014a,b), shifted asymmetric Laplace distributions (Franczak et al., 2014), multivariate power exponential distributions (Dang et al., 2015), multivariate normal inverse Gaussian distributions (Karlis and Santourian, 2009; O’Hagan et al., 2016), and generalized hyperbolic distributions (Browne and McNicholas, 2015; Morris and McNicholas, 2016; Tortora et al., 2016). A comprehensive review of model-based clustering work, up to and including some recent work on non-Gaussian mixtures, is given by McNicholas (2016b).

Unobserved or missing observations are frequently a hindrance in multivariate datasets and so developing mixture models that can accommodate incomplete data is an important issue in model-based clustering. The maximum likelihood and Bayesian approaches are two common imputation paradigms for analyzing data with incomplete observations. Herein, the missing data mechanism is assumed to be missing at random (MAR), as per Rubin (1976) and Little and Rubin (1987), meaning that the probability that a variable is missing for a particular individual depends only on the observed data and not on the value of the missing variable. Note that missing completely at random (MCAR) is a special case of MAR. Under MAR, the missing data mechanisms are ignorable for methods using the maximum likelihood approach.

The maximum likelihood approach to clustering incomplete data has been well studied and is often used, particularly for Gaussian mixture models (e.g., Ghahramani and Jordan, 1994; Lin et al., 2006; Browne et al., 2013). Wang et al. (2004) present a framework maximum likelihood estimation using an expectation-maximization (EM) algorithm (Dempster et al., 1977) to fit a mixture of multivariate t -distributions with arbitrary missing data patterns, which was generalized by Lin et al. (2009) to efficient supervised learning via the

parameter expanded (PX-EM) algorithm (Liu et al., 1998) through two auxiliary indicator matrices. Lin (2014) further develops a family of multivariate- t mixture models with 14 eigen-decomposed scale matrices in the presence of missing data through a computationally flexible EM algorithm by incorporating two auxiliary indicator matrices.

We consider fitting mixtures of generalized hyperbolic distributions (MGHD) and mixtures of multivariate skew- t distributions (MST) with missing information. In each case, an EM algorithm is used for model selection. The chosen formulation of the (multivariate) generalized hyperbolic distribution (GHD) is that used by Browne and McNicholas (2015) and has formulations of several well-known distributions as special cases such as the multivariate skew- t , normal inverse Gaussian, variance-gamma, Laplace, and Gaussian distributions (cf. McNeil et al., 2005). In addition to considering missing data, we develop families of MGHD and MST mixture models, each with 14 parsimonious eigen-decomposed scale matrices corresponding to the famous Gaussian parsimonious clustering models (GPCMs) of Banfield and Raftery (1993) and Celeux and Govaert (1995).

2 Background

2.1 Generalized Inverse Gaussian Distribution

The random variable $W \in \mathbb{R}^+$ is said to have a generalized inverse Gaussian (GIG) distribution, introduced by (Good, 1953), with parameters λ , χ , and ψ if its probability density function is given by

$$f_{\text{GIG}}(w \mid \lambda, \chi, \psi) = \frac{(\psi/\chi)^{\lambda/2} w^{\lambda-1}}{2K_{\lambda}(\sqrt{\psi\chi})} \exp \left\{ -\frac{\psi w + \chi/w}{2} \right\}, \quad (1)$$

where $\psi, \chi \in \mathbb{R}^+, \lambda \in \mathbb{R}$, and K_{λ} is the modified Bessel function of the third kind with index λ . Herein, we write $W \sim \text{GIG}(\lambda, \chi, \psi)$ to indicate that a random variable W has the GIG density as parameterized in (1). The GIG distribution has some attractive properties (Barndorff-Nielsen and Halgreen, 1977; Blæsild, 1978; Halgreen, 1979; Jørgensen, 1982), including the tractability of the expectations:

$$\mathbb{E}[W^{\alpha}] = \left(\frac{\chi}{\psi} \right)^{\alpha/2} \frac{K_{\lambda+\alpha}(\sqrt{\psi\chi})}{K_{\lambda}(\sqrt{\psi\chi})}, \quad (2)$$

for $\alpha \in \mathbb{R}$, and

$$\mathbb{E}[\log W] = \log \left(\sqrt{\frac{\chi}{\psi}} \right) + \frac{\partial}{\partial \lambda} \log(K_\lambda(\sqrt{\psi\chi})). \quad (3)$$

Specifically, for $\alpha = 1$ and $\alpha = -1$, we have

$$\begin{aligned} \mathbb{E}[W] &= \sqrt{\frac{\chi}{\psi}} \frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})}, \\ \mathbb{E}[1/W] &= \sqrt{\frac{\psi}{\chi}} \frac{K_{\lambda-1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} = \sqrt{\frac{\psi}{\chi}} \frac{K_{\lambda+1}(\sqrt{\psi\chi})}{K_\lambda(\sqrt{\psi\chi})} - \frac{2\lambda}{\chi}. \end{aligned}$$

Browne and McNicholas (2015) introduce another parameterization of the GIG distribution by setting $\omega = \sqrt{\psi\chi}$ and $\eta = \sqrt{\chi/\psi}$. Write $W \sim \mathcal{I}(\lambda, \eta, \omega)$; its density is given by

$$f_{\mathcal{I}}(w \mid \lambda, \eta, \omega) = \frac{(w/\eta)^{\lambda-1}}{2\eta K_\lambda(\omega)} \exp \left\{ -\frac{\omega}{2} \left(\frac{w}{\eta} + \frac{\eta}{w} \right) \right\} \quad (4)$$

for $w > 0$, where $\eta \in \mathbb{R}^+$ is a scale parameter and $\omega \in \mathbb{R}^+$ is a concentration parameter. These two parameterizations of the GIG distribution are important ingredients for building the generalized hyperbolic distribution presented later.

2.2 Generalized Hyperbolic Distribution

Several alternative parameterizations of the GHD have appeared in the literature, e.g., Barndorff-Nielsen and Blæsild (1981), McNeil et al. (2005), and Browne and McNicholas (2015). Barndorff-Nielsen (1977) introduces the generalized hyperbolic distribution (GHD) to model the distribution of the sand grain sizes and subsequent reports described its statistical properties (e.g., Barndorff-Nielsen, 1978; Barndorff-Nielsen and Blæsild, 1981). However, under this standard parameterization, the parameters of the mixing distribution are not invariant by affine transformations. An important innovation was made by McNeil et al. (2005), who gave a new parameterization of the GHD. Under this new parameterization, the linear transformation of GHD remains in the same sub-family characterized by the parameters of the mixing distribution. However, there is an identifiability issue arising under this parameterization. To solve this problem, Browne and McNicholas (2015) give an alternative parameterization.

Following McNeil et al. (2005), a $p \times 1$ random vector \mathbf{X} is said to follow a generalized hyperbolic distribution with index parameter λ , concentration parameters χ and ψ , location

vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\alpha}$, denoted by $\mathbf{X} \sim \text{GH}_p(\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$, if it can be represented by

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\alpha} + \sqrt{W}\mathbf{U}, \quad \mathbf{U} \perp W \quad (5)$$

where $W \sim \text{GIG}(\lambda, \chi, \psi)$, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, the symbol \perp indicates independence, and it follows that $\mathbf{X} | w \sim \mathcal{N}(\boldsymbol{\mu} + w\boldsymbol{\alpha}, w\boldsymbol{\Sigma})$. So, the density of the generalized hyperbolic random vector \mathbf{X} is given by

$$f(\mathbf{x} | \boldsymbol{\vartheta}) = \left[\frac{\chi + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma})}{\psi + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}} \right]^{\frac{\lambda-p/2}{2}} \frac{(\psi/\chi)^{\lambda/2} K_{\lambda-p/2} \left(\sqrt{(\chi + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}))(\psi + \boldsymbol{\alpha}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha})} \right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\sqrt{\chi\psi}) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\alpha}\}}, \quad (6)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}) = (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is the squared Mahalanobis distance between \mathbf{x} and $\boldsymbol{\mu}$, K_λ is the modified Bessel function of the third kind with index λ , and $\boldsymbol{\vartheta} = (\lambda, \chi, \psi, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\alpha})$ denotes the model parameters. The mean and covariance matrix of \mathbf{X} are

$$\mathbb{E}(\mathbf{X}) = \boldsymbol{\mu} + \mathbb{E}(W)\boldsymbol{\alpha} \quad \text{and} \quad \text{Var}(\mathbf{X}) = \mathbb{E}(W)\boldsymbol{\Sigma} + \text{Var}(W)\boldsymbol{\alpha}\boldsymbol{\alpha}^\top, \quad (7)$$

respectively, where $\mathbb{E}(W)$ and $\text{Var}(W)$ are the mean and variance of the random variable W , respectively.

Note that, in this parameterization, we need to hold $|\boldsymbol{\Sigma}| = 1$ to ensure identifiability. Using $|\boldsymbol{\Sigma}| = 1$ solves the identifiability problem but would be prohibitively restrictive for model-based clustering and classification applications. Hence, Browne and McNicholas (2015) develop a new parameterization of the GHD with index parameter λ , concentration parameter ω , location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\beta} = \eta\boldsymbol{\alpha}$, denoted by $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Note that $\eta = 1$. This formulation is given by

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\beta} + \sqrt{W}\mathbf{U}, \quad (8)$$

where $W \sim \mathcal{I}(\lambda, 1, \omega)$ and $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Under this parameterization, the density of the generalized hyperbolic random vector \mathbf{X} is

$$f(\mathbf{x} | \boldsymbol{\vartheta}) = \left[\frac{\omega + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma})}{\omega + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}} \right]^{\frac{\lambda-p/2}{2}} \frac{K_{\lambda-p/2} \left(\sqrt{(\omega + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}))(\omega + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta})} \right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\omega) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\}}, \quad (9)$$

where $\delta(\mathbf{x}, \boldsymbol{\mu} \mid \boldsymbol{\Sigma})$ and $K_{\lambda-p/2}$ are as described earlier. We use this parameterization when we describe parameter estimation (cf. Section 3).

The following result shows an appealing closure property of the generalized hyperbolic distribution under affine transformation and conditioning as well as the formation of marginal distributions, which is useful for developing new methods presented later. Suppose that \mathbf{X} is a p -dimensional random vector having a generalized hyperbolic distribution as in (9), i.e., $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Assume that \mathbf{X} is partitioned as $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where \mathbf{X}_1 takes values in \mathbb{R}^{d_1} and \mathbf{X}_2 in $\mathbb{R}^{d_2} = \mathbb{R}^{p-d_1}$, with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix}, \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where \mathbf{X} , $\boldsymbol{\mu}$, and $\boldsymbol{\beta}$ have similar partitions. Furthermore, $\boldsymbol{\Sigma}_{11}$ is $d_1 \times d_1$ and $\boldsymbol{\Sigma}_{22}$ is $d_2 \times d_2$.

Proposition 1. *Affine transformation of the generalized hyperbolic distribution. If $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ and $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$ where $\mathbf{B} \in \mathbb{R}^{k \times p}$ and $\mathbf{b} \in \mathbb{R}^k$, then*

$$\mathbf{Y} \sim \text{GHD}_k(\lambda, \omega, \mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top, \mathbf{B}\boldsymbol{\beta}), \quad (10)$$

Proof. The result follows by substituting (8) into $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. \square

Proposition 2. *The marginal distribution of \mathbf{X}_1 is a generalized hyperbolic distribution as in (9) with index parameter λ , concentration parameter ω , location vector $\boldsymbol{\mu}_1$, dispersion matrix $\boldsymbol{\Sigma}_{11}$, and skewness vector $\boldsymbol{\beta}_1$, i.e., $\mathbf{X}_1 \sim \text{GHD}_{d_1}(\lambda, \omega, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\beta}_1)$.*

Proof. The result follows by applying Proposition 1 and choosing $\mathbf{B} = [\mathbf{I}_{d_1}, \mathbf{0}]$ and $\mathbf{b} = \mathbf{0}$. The parameters λ, ω inherited from the mixing distribution $W \sim \mathcal{I}(\lambda, \eta = 1, \omega)$ remain the same under the affine transformation and marginal distribution. \square

Proposition 3. *The conditional distribution of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$ is a generalized hyperbolic distribution as in (6), i.e., $\mathbf{X}_2 \mid \mathbf{X}_1 = \mathbf{x}_1 \sim \text{GH}_{d_2}(\lambda_{2|1}, \chi_{2|1}, \psi_{2|1}, \boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}, \boldsymbol{\beta}_{2|1})$, where*

$$\begin{aligned} \lambda_{2|1} &= \lambda - \frac{d_1}{2}, & \chi_{2|1} &= \omega + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ \psi_{2|1} &= \omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1, & \boldsymbol{\mu}_{2|1} &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ \boldsymbol{\Sigma}_{2|1} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}, & \boldsymbol{\beta}_{2|1} &= \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1. \end{aligned}$$

The proof of Proposition 3 is given in Appendix B.

2.3 The Multivariate Skew- t Distribution

There are several alternative formulations of multivariate skew- t distributions appearing in the literature (e.g., Branco and Dey, 2001; Sahu, Dey, and Branco, 2003; Murray, Browne, and McNicholas, 2014a; Lin, Wu, McLachlan, and Lee, 2014; Lee and McLachlan, 2014). Lin and Lin (2011) develop a mixture of multivariate skew- t distributions incomplete data using the formulation of Sahu et al. (2003). Herein, the formulation of the multivariate skew- t distribution arising from the generalized hyperbolic distribution is used. This formulation of the multivariate skew- t distribution has been used by Murray et al. (2014a) to develop a mixture of skew- t factor analyzers model.

Following McNeil et al. (2005), a $p \times 1$ random vector \mathbf{X} is said to follow a multivariate skew- t distribution with degree of freedom parameter v , location vector $\boldsymbol{\mu}$, dispersion matrix $\boldsymbol{\Sigma}$, and skewness vector $\boldsymbol{\beta}$, denoted by $\mathbf{X} \sim \text{ST}_p(v, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$, if it can be represented by

$$\mathbf{X} = \boldsymbol{\mu} + W\boldsymbol{\beta} + \sqrt{W}\mathbf{U}, \quad (11)$$

where $W \sim \text{IG}(v/2, v/2)$, $\mathbf{U} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$, with $\text{IG}(\cdot)$ denoting the inverse Gamma distribution. It follows that $\mathbf{X} | w \sim \mathcal{N}(\boldsymbol{\mu} + w\boldsymbol{\beta}, w\boldsymbol{\Sigma})$ and the pdf of the multivariate skew- t random vector \mathbf{X} is given by

$$f(\mathbf{x} | \boldsymbol{\vartheta}) = \left[\frac{v + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma})}{\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}} \right]^{\frac{-v-p}{4}} \frac{v^{v/2} K_{(-v-p)/2} \left(\sqrt{(v + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}))(\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta})} \right)}{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} \Gamma(v/2) 2^{v/2-1} \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\}}. \quad (12)$$

This formulation of the multivariate skew- t distribution can be obtained as a special case of the generalized hyperbolic distribution by setting $\lambda = -v/2$ and $\chi = v$, and letting $\psi \rightarrow 0$. Similarly, this formulation of the multivariate skew- t distribution has a closed form under affine transformation and conditioning, and the formation of marginal distributions, which is useful for developing new methods presented later. Suppose that \mathbf{X} is a p -dimensional random vector having the multivariate skew- t distribution as in (12), i.e., $\mathbf{X} \sim \text{ST}_p(v, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$. Assume that \mathbf{X} is partitioned as $\mathbf{X} = (\mathbf{X}_1^\top, \mathbf{X}_2^\top)^\top$, where \mathbf{X}_1 takes values in \mathbb{R}^{d_1} and \mathbf{X}_2 in $\mathbb{R}^{d_2} = \mathbb{R}^{p-d_1}$, with

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix},$$

where \mathbf{X} , $\boldsymbol{\mu}$, and $\boldsymbol{\beta}$ have similar partitions. Furthermore, $\boldsymbol{\Sigma}_{11}$ is $d_1 \times d_1$ and $\boldsymbol{\Sigma}_{22}$ is $d_2 \times d_2$.

Proposition 4. *Affine transformation of the multivariate skew-t distribution. If $\mathbf{X} \sim ST_p(v, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ and $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$, where $\mathbf{B} \in \mathbb{R}^{k \times p}$ and $\mathbf{b} \in \mathbb{R}^p$, then*

$$\mathbf{Y} \sim ST_k(v, \mathbf{B}\boldsymbol{\mu} + \mathbf{b}, \mathbf{B}\boldsymbol{\Sigma}\mathbf{B}^\top, \mathbf{B}\boldsymbol{\beta}). \quad (13)$$

Proof. The proof follows easily by substituting (11) into $\mathbf{Y} = \mathbf{B}\mathbf{X} + \mathbf{b}$. \square

Proposition 5. *The marginal distribution of \mathbf{X}_1 is a multivariate skew-t distribution as in (12) with degree of freedom parameter v , location vector $\boldsymbol{\mu}_1$, dispersion matrix $\boldsymbol{\Sigma}_{11}$, and skewness vector $\boldsymbol{\beta}_1$, i.e., $\mathbf{X}_1 \sim ST_{d_1}(v, \boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}, \boldsymbol{\beta}_1)$.*

Proof. The proof follows easily by applying Proposition 4 and choosing $\mathbf{B} = [\mathbf{I}_{d_1}, \mathbf{0}]$ and $\mathbf{b} = \mathbf{0}$. The degree of freedom parameter v inherited from the mixing distribution $W \sim \text{IG}(v/2, v/2)$ remains invariant under affine transformation and marginal distribution. \square

Proposition 6. *The conditional distribution of \mathbf{X}_2 given $\mathbf{X}_1 = \mathbf{x}_1$ is a generalized hyperbolic distribution as in (6), i.e., $\mathbf{X}_2 \mid \mathbf{x}_1 \sim GH_{d_2}(\lambda_{2|1}, \chi_{2|1}, \psi_{2|1}, \boldsymbol{\mu}_{2|1}, \boldsymbol{\Sigma}_{2|1}, \boldsymbol{\beta}_{2|1})$, where*

$$\begin{aligned} \lambda_{2|1} &= -(v + d_1)/2, & \chi_{2|1} &= v + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ \psi_{2|1} &= \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1, & \boldsymbol{\mu}_{2|1} &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ \boldsymbol{\Sigma}_{2|1} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}, & \boldsymbol{\beta}_{2|1} &= \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1. \end{aligned}$$

The proof of Proposition 6 is similar to that for Proposition 3, hence is omitted.

3 MGHD with Incomplete Data

Let $\mathbf{X}_1, \dots, \mathbf{X}_n$ be p -dimensional random variables arising from a heterogeneous population with G disjoint MGHD subpopulations. That is, each \mathbf{X}_i has the density

$$f_{\text{MGHD}}(\mathbf{x}_i \mid \boldsymbol{\Theta}) = \sum_{g=1}^G \pi_g f_{\text{GHD}}(\mathbf{x}_i \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \quad (14)$$

where $\pi_g > 0$, such that $\sum_{g=1}^G \pi_g = 1$ are the mixing proportions, $\boldsymbol{\Theta}$ denotes the model parameters, and $f_{\text{GHD}}(\mathbf{X}_i \mid \lambda_g, \omega_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g)$ is the GHD density defined in (9).

To apply the MGHD model (14) in the clustering paradigm, introduce $\mathbf{Z}_i = (Z_{i1}, \dots, Z_{iG})^\top$, where $Z_{ig} = 1$ if observation i is in component g and $Z_{ig} = 0$ otherwise. We have $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \dots, \pi_G)$, i.e., \mathbf{Z}_i follows a multinomial distribution with one trial and cell probabilities π_1, \dots, π_G .

A three-level hierarchical representation of the MGHD model (14) can be expressed by

$$\begin{aligned}\mathbf{X}_i \mid (w_{ig}, Z_{ig} = 1) &\sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g), \\ W_{ig} \mid (Z_{ig} = 1) &\sim \mathcal{I}(\lambda_g, \eta = 1, \omega_g), \\ \mathbf{Z}_i &\sim \mathcal{M}(1; \pi_1, \dots, \pi_G).\end{aligned}\tag{15}$$

The complete-data consist of the observed \mathbf{x}_i together with the missing group membership z_{ig} and the latent w_{ig} , for $i = 1, \dots, n$ and $g = 1, \dots, G$, and the complete-data log-likelihood is given by

$$l_c(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} [\log(\pi_g) + \log(\phi(\mathbf{x}_i \mid \boldsymbol{\mu}_g + w_{ig}\boldsymbol{\beta}_g, w_{ig}\boldsymbol{\Sigma}_g)) + \log(h(w_{ig} \mid \lambda_g, \omega_g))]. \tag{16}$$

Browne and McNicholas (2015) present an EM algorithm for parameter estimation with the MGHD when there is no missing data in $\mathbf{x}_1, \dots, \mathbf{x}_n$. We are interested in parameter estimation for the MGHD model (14) when $\mathbf{x}_1, \dots, \mathbf{x}_n$ are partially observed with arbitrary missing patterns. The missing data mechanism is assumed to be MAR. Assume now that we split \mathbf{x}_i into two components, \mathbf{x}_i^o and \mathbf{x}_i^m that denote the observed and missing components of \mathbf{x}_i , respectively. In general, each data vector \mathbf{x}_i may have a different pattern of missing features, i.e., $\mathbf{x}_i = (\mathbf{x}_i^{o\top}, \mathbf{x}_i^{m\top})^\top$, but can be simplified for the sake of clarity.

For each $\mathbf{x}_i = (\mathbf{x}_i^{o\top}, \mathbf{x}_i^{m\top})^\top$, partition the vector mean $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_{g,i}^{o\top}, \boldsymbol{\mu}_{g,i}^{m\top})^\top$, where $\boldsymbol{\mu}_{g,i}^o$ and $\boldsymbol{\mu}_{g,i}^m$ denote the sub-vectors of $\boldsymbol{\mu}_g$ matching the observed and missing components of \mathbf{x}_i , respectively. Similarly, the skewness vector is $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g,i}^{o\top}, \boldsymbol{\beta}_{g,i}^{m\top})^\top$ and the covariance matrix $\boldsymbol{\Sigma}_g$ as

$$\boldsymbol{\Sigma}_g = \begin{pmatrix} \boldsymbol{\Sigma}_{g,i}^{oo} & \boldsymbol{\Sigma}_{g,i}^{om} \\ \boldsymbol{\Sigma}_{g,i}^{mo} & \boldsymbol{\Sigma}_{g,i}^{mm} \end{pmatrix} \text{ and } \boldsymbol{\Sigma}_g^{-1} = \begin{pmatrix} \boldsymbol{\Sigma}_{g,i}^{-1,oo} & \boldsymbol{\Sigma}_{g,i}^{-1,om} \\ \boldsymbol{\Sigma}_{g,i}^{-1,mo} & \boldsymbol{\Sigma}_{g,i}^{-1,mm} \end{pmatrix}, \tag{17}$$

correspond to $\mathbf{x}_i = (\mathbf{x}_i^{o\top}, \mathbf{x}_i^{m\top})^\top$. As a result, in addition to the observed \mathbf{x}_i^o , the missing group membership z_{ig} , and the latent variable w_{ig} , the complete-data also include the missing data \mathbf{x}_i^m . In the framework of the EM algorithm, the missing data \mathbf{x}_i^m are considered to be random

variables that are updated in each iteration. Hence, the complete-data log-likelihood (16) is rewritten as

$$l_c(\Theta) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} [\log \pi_g + \log \phi(\mathbf{x}_i^o, \mathbf{x}_i^m \mid \boldsymbol{\mu}_g + w_{ig} \boldsymbol{\beta}_g, w_{ig} \boldsymbol{\Sigma}_g) + \log h_{\mathcal{I}}(w_{ig} \mid \lambda_g, \omega_g)].$$

Given (15), we establish the following:

- The marginal distribution of \mathbf{X}_i^o given is

$$\mathbf{X}_i^o \sim \sum_{g=1}^G \pi_g f_{\text{GHD}, p_i^o}(\lambda_g, \omega_g, \boldsymbol{\mu}_{g,i}^o, \boldsymbol{\Sigma}_{g,i}^{oo}, \boldsymbol{\beta}_{g,i}^o),$$

where p_i^o is the dimension corresponding to the observed component \mathbf{x}_i^o , which should be exactly written as $p_i^{o_i}$ but here is simplified.

- The conditional distribution of \mathbf{X}_i^m given \mathbf{x}_i^o and $Z_{ig} = 1$, according to Proposition 3, is

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, Z_{ig} = 1 \sim \text{GH}_{p-p_i^o} \left(\lambda_{g,i}^{m|o}, \chi_{g,i}^{m|o}, \psi_{g,i}^{m|o}, \boldsymbol{\mu}_{g,i}^{m|o}, \boldsymbol{\Sigma}_{g,i}^{m|o}, \boldsymbol{\beta}_{g,i}^{m|o} \right), \quad (18)$$

where

$$\begin{aligned} \lambda_{g,i}^{m|o} &= \lambda_g - \frac{p_i^o}{2}, & \chi_{g,i}^{m|o} &= \omega_g + (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o), \\ \psi_{g,i}^{m|o} &= \omega_g + \boldsymbol{\beta}_{g,i}^{o\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o, & \boldsymbol{\mu}_{g,i}^{m|o} &= \boldsymbol{\mu}_g^m + \boldsymbol{\Sigma}_{g,i}^{om\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o), \\ \boldsymbol{\Sigma}_{g,i}^{m|o} &= \boldsymbol{\Sigma}_{g,i}^{mm} - \boldsymbol{\Sigma}_{g,i}^{om\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\Sigma}_{g,i}^{om}, & \boldsymbol{\beta}_{g,i}^{m|o} &= \boldsymbol{\beta}_{g,i}^m - \boldsymbol{\Sigma}_{g,i}^{om\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o. \end{aligned}$$

- The conditional distribution of \mathbf{X}_i^m given \mathbf{x}_i^o , w_{ig} , and $Z_{ig} = 1$ is

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, w_{ig}, Z_{ig} = 1 \sim \mathcal{N}_{p-p_i^o}(\boldsymbol{\mu}_{g,i}^{m|o} + w_{ig} \boldsymbol{\beta}_{g,i}^{m|o}, w_{ig} \boldsymbol{\Sigma}_{g,i}^{m|o}). \quad (19)$$

- The conditional distribution of W_i given \mathbf{x}_i^o and $Z_{ig} = 1$ is

$$W_{ig} \mid \mathbf{x}_i^o, Z_{ig} = 1 \sim \text{GIG} \left(\omega_g + \boldsymbol{\beta}_{g,i}^{o\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o, \omega_g + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^o \mid \boldsymbol{\Sigma}_{g,i}^{oo}), \lambda_g - \frac{p_i^o}{2} \right). \quad (20)$$

After a little algebra, we get the complete data log-likelihood function is

$$\begin{aligned}
l_c(\Theta) = & \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log \pi_g + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[-\frac{p}{2} \log(2\pi) - \frac{p}{2} \log w_{ig} + \frac{1}{2} \log |\Sigma_g^{-1}| \right] \\
& - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\Sigma_g^{-1} z_{ig} \frac{1}{w_{ig}} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \\ (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o) & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\
& + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\Sigma_g^{-1} z_{ig} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^o \\ \boldsymbol{\beta}_{g,i}^m \end{pmatrix} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\
& + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\Sigma_g^{-1} z_{ig} \begin{pmatrix} \mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o \\ \mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^{o\top} & \boldsymbol{\beta}_{g,i}^{m\top} \end{pmatrix} \right) - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} w_{ig} \boldsymbol{\beta}_{g,i}^\top \Sigma_g^{-1} \boldsymbol{\beta}_{g,i} \\
& + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[(\lambda_g - 1) \log w_{ig} - \log(2K_{\lambda_g}(\omega_g)) - \frac{\omega_g}{2} \left(w_{ig} + \frac{1}{w_{ig}} \right) \right].
\end{aligned} \tag{21}$$

On the k th iteration of the E-step, the expected value of the complete data log-likelihood is computed given the observed data $\mathbf{x}_1^o, \dots, \mathbf{x}_n^o$ and the current parameter updates $\Theta^{(k)}$. That is, we need to compute $\mathbb{E}(Z_{ig} \mid \mathbf{x}_i^o; \Theta^{(k)})$, $\mathbb{E}(W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \Theta^{(k)})$, $\mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \Theta^{(k)})$, $\mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \Theta^{(k)})$, $\mathbb{E}(\mathbf{X}_i^m \mid \mathbf{x}_i^o, z_{ig} = 1, w_i; \Theta^{(k)})$, and $\mathbb{E}(\mathbf{X}_i^m \mathbf{X}_i^{m\top} \mid \mathbf{x}_i^o, z_{ig} = 1, w_i; \Theta^{(k)})$.

First, let $\hat{z}_{ig}^{(k)}$ denote the *a posteriori* probability that i -th observation belongs to the g -th component of the mixture, based on the observed data:

$$\hat{z}_{ig}^{(k)} := \mathbb{E}(Z_{ig} \mid \mathbf{x}_i^o, \Theta^{(k)}) = \frac{\pi_g^{(k)} f_{\text{GHD}, p_i^o}(\mathbf{x}_i^o; \lambda_g^{(k)}, \omega_g^{(k)}, \boldsymbol{\mu}_{g,i}^{o(k)}, \boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)}, \boldsymbol{\beta}_{g,i}^{o(k)})}{\sum_{l=1}^G \pi_l^{(k)} f_{\text{GHD}, p_i^o}(\mathbf{x}_i^o; \lambda_l^{(k)}, \omega_l^{(k)}, \boldsymbol{\mu}_{l,i}^{o(k)}, \boldsymbol{\Sigma}_{l,i}^{\text{oo}(k)}, \boldsymbol{\beta}_{l,i}^{o(k)})}.$$

Given (2), (3), and (20), we have the following expectations as to the latent variable W :

$$\begin{aligned}
a_{ig}^{(k)} := \mathbb{E}(W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \Theta^{(k)}) = & \sqrt{\frac{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)})}{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}} \\
& \times \frac{K_{\lambda_g^{(k)} - \frac{p_i^0}{2} + 1} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)})) (\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{\lambda_g^{(k)} - \frac{p_i^0}{2}} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)})) (\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{\text{oo}(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)},
\end{aligned}$$

$$\begin{aligned}
b_{ig}^{(k)} &:= \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) \\
&= -\frac{2\lambda_g^{(k)} - p_i^o}{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})} + \sqrt{\frac{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}} \\
&\quad \times \frac{K_{\lambda_g^{(k)} - \frac{p_i^o}{2} + 1} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{\lambda_g^{(k)} - \frac{p_i^o}{2}} \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}, \\
c_{ig}^{(k)} &:= \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \log \left(\sqrt{\frac{\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}} \right) \\
&\quad + \frac{\partial}{\partial t} \log \left\{ K_t \left(\sqrt{(\omega_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\omega_g^{(k)} + \boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right) \right\} \Big|_{t=(\lambda_g^{(k)} - \frac{p_i^o}{2})}.
\end{aligned}$$

For convenience, we use the following notation analogous to Browne and McNicholas (2015):

$$n_g^{(k)} = \sum_{i=1}^n \hat{z}_{ig}^{(k)}, \bar{a}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} a_{ig}^{(k)}, \bar{b}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} b_{ig}^{(k)}, \text{ and } \bar{c}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} c_{ig}^{(k)}.$$

For the actual missing data \mathbf{X}^m , we will also need the following expectations:

$$\begin{aligned}
\hat{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}(\mathbf{X}_i^m \mid \mathbf{x}_i^o, Z_{ig} = 1) = \boldsymbol{\mu}_{g,i}^{m|o(k)} + a_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)}, \\
\tilde{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}((1/W_i) \mathbf{X}_i^m \mid \mathbf{x}_i^o, Z_{ig} = 1) = b_{ig}^{(k)} \boldsymbol{\mu}_{g,i}^{m|o(k)} + \boldsymbol{\beta}_{g,i}^{m|o(k)}, \\
\tilde{\tilde{\mathbf{x}}}_{ig}^{m(k)} &:= \mathbb{E}((1/W_i) \mathbf{X}_i^m \mathbf{X}_i^{m\top} \mid \mathbf{x}_i^o, Z_{ig} = 1) = \boldsymbol{\Sigma}_{g,i}^{m|o(k)} + b_{ig}^{(k)} \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top \\
&\quad + \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top + \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top + a_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top.
\end{aligned}$$

On the k -th iteration of the M-step, the expected value of the complete data log-likelihood is maximized to get the updates for the parameter estimates as follows:

$$\begin{aligned}
\pi_g^{(k+1)} &= \frac{n_g^{(k)}}{n}, \\
\boldsymbol{\mu}_g^{(k+1)} &= \frac{1}{\sum_{i=1}^n \hat{z}_{ig}^{(k)} (\bar{a}_g^{(k)} b_{ig}^{(k)} - 1)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} \left(\frac{(\bar{a}_g^{(k)} b_{ig}^{(k)} - 1) \mathbf{x}_i^o}{\bar{a}_g^{(k)} \tilde{\mathbf{x}}_{ig}^{m(k)} - \hat{\mathbf{x}}_{ig}^{m(k)}} \right), \\
\boldsymbol{\beta}_g^{(k+1)} &= \frac{1}{\sum_{i=1}^n \hat{z}_{ig}^{(k)} (\bar{a}_g^{(k)} b_{ig}^{(k)} - 1)} \sum_{i=1}^n \hat{z}_{ig}^{(k)} \left(\frac{(\bar{b}_g^{(k)} - b_{ig}^{(k)}) \mathbf{x}_i^o}{\bar{b}_g^{(k)} \tilde{\mathbf{x}}_{ig}^{m(k)} - \tilde{\tilde{\mathbf{x}}}_{ig}^{m(k)}} \right), \\
\boldsymbol{\Sigma}_g^{(k+1)} &= \frac{1}{n_g^{(k)}} \sum_{i=1}^n \hat{z}_{ig}^{(k)} \boldsymbol{\Sigma}_{ig}^{(k+1)} - (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)}) \boldsymbol{\beta}_g^{(k+1)\top} - \boldsymbol{\beta}_g^{(k+1)} (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)})^\top + \bar{a}_g^{(k)} \boldsymbol{\beta}_g^{(k+1)} \boldsymbol{\beta}_g^{(k+1)\top},
\end{aligned}$$

where

$$\bar{\mathbf{x}}_g = \frac{1}{n_g^{(k+1)}} \sum_{i=1}^n \hat{z}_{ig}^{(k+1)} \begin{pmatrix} \mathbf{x}_i^o \\ \hat{\mathbf{x}}_{ig}^{m(k+1)} \end{pmatrix},$$

$$\Sigma_{ig}^{(k+1)} = \begin{pmatrix} b_{ig}^{(k)}(\mathbf{x}_i^o - \boldsymbol{\mu}_g^{o(k+1)})(\mathbf{x}_i^o - \boldsymbol{\mu}_g^{o(k+1)})^\top & (\mathbf{x}_i^o - \hat{\boldsymbol{\mu}}_g^{o(k+1)})(\tilde{\mathbf{x}}_{ig}^{m(k)} - b_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{m(k+1)})^\top \\ (\tilde{\mathbf{x}}_{ig}^{m(k)} - b_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{m(k+1)})(\mathbf{x}_i^o - \boldsymbol{\mu}_g^{o(k+1)})^\top & \mathbf{k}_{ig}^{m(k+1)} \end{pmatrix},$$

where

$$\mathbf{k}_{ig}^{m(k+1)} = \tilde{\mathbf{x}}_{ig}^{m(k)} - \tilde{\mathbf{x}}_{ig}^{m(k)}\hat{\boldsymbol{\mu}}_g^{m(k+1)\top} - \hat{\boldsymbol{\mu}}_g^{m(k+1)}\tilde{\mathbf{x}}_i^{m(k)\top} + b_{ig}^{(k)}\hat{\boldsymbol{\mu}}_g^{m(k+1)}\hat{\boldsymbol{\mu}}_g^{m(k+1)\top}.$$

Finally, the estimates of $\lambda_g^{(k+1)}$ and $\omega_g^{(k+1)}$ are given as solutions to maximize the function

$$q_g(\lambda_g, \omega_g) = -\log(K_{\lambda_g}(\omega_g)) + (\lambda_g - 1)\bar{c}_g - \frac{\omega_g}{2}(\bar{a}_g + \bar{b}_g),$$

and the associated updates are

$$\lambda_g^{(k+1)} = \bar{c}_g^{(k)}\lambda_g^{(k)} \left[\frac{\partial}{\partial \lambda_g^{(k)}} \log \left(K_{\lambda_g^{(k)}}(\omega_g^{(k)}) \right) \right]^{-1},$$

$$\omega_g^{(k+1)} = \omega_g^{(k)} - \left[\frac{\partial}{\partial \omega_g^{(k)}} q_g(\lambda_g^{(k+1)}, \omega_g^{(k)}) \right] \left[\frac{\partial^2}{\partial \omega_g^{2(k)}} q_g(\lambda_g^{(k+1)}, \omega_g^{(k)}) \right]^{-1}.$$

Details on the MST with incomplete data are analogous to the MGHD with incomplete data and are given in Appendix C.

4 Notes on Implementation

4.1 Initial values

It is well known that the EM algorithm can be heavily dependent on the initial values; indeed, good initial values of parameter estimates may speed up convergence. In this study, the following procedure for automatically generating initial values is used, unless otherwise specified.

- Fill in the missing values based on the mean imputation method.
- Perform k -means clustering and use the resulting clustering membership to initialize

the *a posteriori* probability \hat{z}_{ig}^0 . Accordingly, the initial values for the model parameters are then given by:

$$\hat{\pi}_g^{(0)} = \frac{\sum_{i=1}^n \hat{z}_{ig}^0}{n}, \quad \hat{\boldsymbol{\mu}}_g^{(0)} = \frac{\sum_{i=1}^n \hat{z}_{ig}^0 \mathbf{x}_i}{\sum_{i=1}^n \hat{z}_{ig}^0}, \quad \hat{\boldsymbol{\Sigma}}_g^{(0)} = \frac{\sum_{i=1}^n \hat{z}_{ig}^0 (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(0)}) (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_g^{(0)})^\top}{\sum_{i=1}^n \hat{z}_{ig}^0}.$$

- Set the skewness parameter $\boldsymbol{\beta}_g^{(0)}$ to be close to zero for symmetric data.
- When applicable, we set $\omega_g^{(0)} = 1$ and $\lambda_g^{(0)} = -0.5$ for the index and concentration parameters and set $v_g^{(0)} = 50$ for the near-normality assumption.

To enhance the computational efficiency of the EM algorithm, we update the parameters per missing pattern instead of per individual. We suggest rearranging \mathbf{X} according to unique patterns of the missing data. The procedure can be implemented as follows:

- Build a binary n by p indicator matrix $\mathbf{R} = [r_{ij}]$, with each entry $r_{ij} = 1$ if \mathbf{X}_{ij} is missing and $r_{ij} = 0$ otherwise;
- Find all unique missing patterns; and
- Update parameters per missing pattern instead of per individual.

4.2 Model Selection and Stopping Criterion

In general, the number of mixture components G is not known *a priori*, and needs to be estimated from the data. Two widely used model selection techniques are the Bayesian information criterion (BIC; Schwarz, 1978) and the integrated completed likelihood (ICL; Biernacki et al., 2000), which are given respectively by

$$\text{BIC} = -2l(\mathbf{x}, \hat{\boldsymbol{\Theta}}) + \rho \log(n) \quad \text{and} \quad \text{ICL} \approx \text{BIC} + 2 \sum_{i=1}^n \sum_{g=1}^G \text{MAP} \{ \hat{z}_{ig} \} \log(\hat{z}_{ig}),$$

where $l(\hat{\boldsymbol{\Theta}})$ is the maximized log-likelihood evaluated at the maximum likelihood estimate $\hat{\boldsymbol{\Theta}}$, ρ is the number of free parameters, n is the number of observations, \hat{z}_{ig} represents the estimated *a posteriori* probability that \mathbf{x}_i arises from the g th component, and MAP denotes the maximum *a posteriori* probability such that $\text{MAP} \{ \hat{z}_{ig} \} = 1$ if $\max_g \{ \hat{z}_{ig} \}$ occurs in the

g th component and $\text{MAP} \{\hat{z}_{ig}\} = 0$ otherwise. The bigger the BIC or ICL value, the better the fitted model.

The EM algorithm can be stopped iterations after the maximum number of iterations, or when the Aitken stopping criterion (Aitken, 1926) is satisfied. The Aitken acceleration at iteration k is

$$a^{(k)} = \frac{l^{(k+1)} - l^{(k)}}{l^{(k)} - l^{(k-1)}},$$

where $l^{(k)}$ is the log-likelihood at iterations k . This yields an asymptotic estimate of the log-likelihood at iteration $k + 1$:

$$l_{\infty}^{(k+1)} = l^{(k)} + \frac{1}{1 - a^{(k)}}(l^{(k+1)} - l^{(k)})$$

(Böhning et al., 1994), and the EM algorithm is stopped when $l_{\infty}^{(k+1)} - l^{(k)} < \epsilon$ (Lindsay, 1995; McNicholas et al., 2010).

5 Numerical Examples

Studies based on both simulated and real datasets are used to compare the clustering performance of the proposed approach. The simulated datasets are each two-component mixtures: a mixture of Gaussian distributions (GMM) with a general VEE covariance structure, a mixture of skew-t distributions (MST) with a diagonal VEI covariance structure, and a mixture of generalized hyperbolic distributions (MGHD) with a general VEE covariance structure. The GMM datasets are generated via the R function `rmvnorm` from the `mvtnorm` package for R, and the MST and MGHD datasets are generated using R code based on the stochastic representations in (11) and (8), respectively.

For each mixture component, $n_g = 200$ two-dimensional vectors \mathbf{x}_i are generated. The presumed parameters of Σ_g ($g = 1, 2$) for the VEE and VEI models are the same as those considered in Celeux and Govaert (1995) and Lin (2014). Each mixture component is centred on a different point giving well-separated and overlapping mixtures. Where applicable, the skewness parameters are $\beta_1 = (1, 1)^\top$ and $\beta_2 = (-1, -1)^\top$, the degrees of freedoms for the MST is $v_1 = v_2 = 7$, and the values of other parameters for the MGHD are $\omega_1 = \omega_2 = 4$ and $\lambda_1 = \lambda_2 = 6$. For each scenario, we create artificially incomplete datasets by removing data through an MAR mechanism from the simulated samples under missing rates r ranging from

5% to 30% while maintaining the condition that each observation has at least one observed attribute. Then our proposed model for incomplete data is compared to the MGHD and MST for complete data once missing data have been ‘filled-in’ with the sample mean of the associated attribute, via the mean imputation method. The misclassification rate δ and the adjusted Rand index (ARI; Hubert and Arabie, 1985) are used to compare predicted classifications with true classes.

5.1 Simulation Studies

The datasets considered in the simulation studies are summarized in Table 1 and plotted in Figure 1. The datasets are overlapping, making this a relatively difficult clustering scenario even when the datasets are complete.

Table 1: Summary of simulated datasets

Dataset	Distribution	Covariance structure	Separation between components
Sim1	MGHD	VEE	well-separated
Sim2	MGHD	VEE	overlapping
Sim3	MST	VEI	well-separated
Sim4	MST	VEI	overlapping
Sim5	GMM	VEE	well-separated
Sim6	GMM	VEE	overlapping

First, we undertook a simulation study similar to those of Celeux and Govaert (1995) and Lin (2014) to investigate the classification performance of the MGHD VEE and MST VEI models with synthetic missing values ($r = 5\%, 30\%$). These two models discussed in this experiment are compared for the six simplest cases among a family of fourteen models. Simulations were run with a total of 80 replicates for the first four simulated datasets. The detailed numerical results are summarized in Tables 2 and 3, including the average misclassification rates with the associated standard deviations in parentheses. The following phenomena are observed, which are similar to results obtained by Lin (2014):

- The average misclassification rate increases as the missing rate rises.
- The overlapping components typically have a higher misclassification rate than well-separated components.

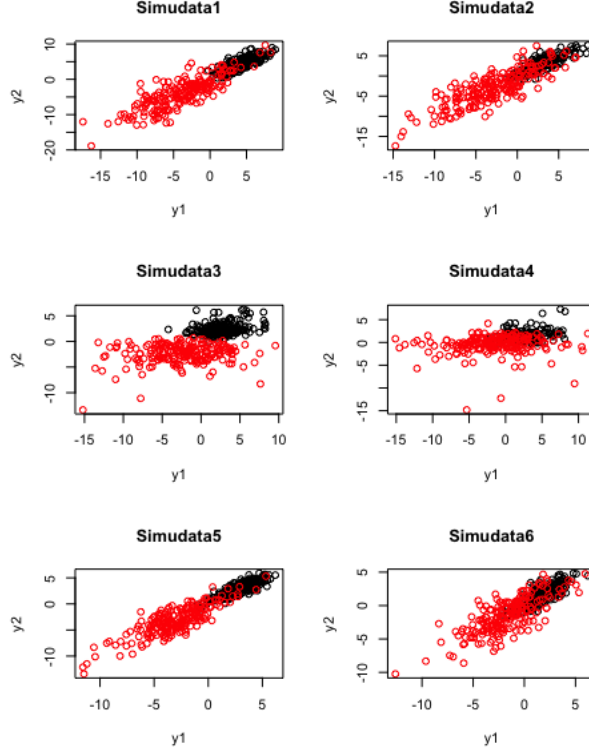


Figure 1: Exemplar scatter plots for simulated datasets.

- The bolded numbers indicate that the best results are generally associated with the true covariance structure.
- The standard deviations increase with missing data rate and the degree of component overlap.

As another illustration, we explore the flexibility of the MGHD model for incomplete data and study the performance of the BIC for model selection. As mentioned in the introduction, the GHD is a flexible distribution with skewness, concentration, and index parameters. The six simulated datasets in Table 1 with missing rates ranging from 5 to 30% were generated under an MAR mechanism with 20 replicates for each dataset. We compute the average misclassification rates, the average ARI, and their associated standard deviations for the parsimonious MGHD and MST models introduced here as well as the mean imputation method under the circumstances of known cluster $G = 2$ and unknown clusters ($G = 1, \dots, 4$). The detailed results are summarized in Tables 7 and 8 (Appendix D). The lowest misclassification rates and the highest ARI are highlighted. From these tables, we observe the following:

Table 2: Misclassification rates and associated standard deviations for each model fitted in Sim1, Sim2, Sim3, and Sim4 when $r = 5\%$.

	EII	VII	EEI	VEI	EEE	VEE
Sim1	0.0670 (0.0421)	0.0538 (0.0125)	0.0717 (0.0529)	0.0546 (0.0134)	0.0557 (0.0210)	0.0531 (0.0131)
Sim2	0.1424 (0.0577)	0.1214 (0.0330)	0.1261 (0.0308)	0.1223 (0.0204)	0.1535 (0.0640)	0.1210 (0.0214)
Sim3	0.0763 (0.0215)	0.0295 (0.0117)	0.0509 (0.0268)	0.0186 (0.0062)	0.0541 (0.0298)	0.0189 (0.0062)
Sim4	0.3050 (0.0631)	0.2019 (0.0458)	0.3316 (0.0752)	0.1907 (0.0353)	0.3306 (0.0582)	0.2425 (0.1004)

Table 3: Misclassification rates and associated standard deviations for each model fitted to Sim1, Sim2, Sim3, and Sim4 when $r = 30\%$.

	EII	VII	EEI	VEI	EEE	VEE
Sim1	0.0784 (0.0416)	0.0646 (0.0192)	0.0853 (0.0498)	0.0671 (0.0203)	0.0692 (0.0297)	0.0601 (0.0305)
Sim2	0.1666 (0.0607)	0.1520 (0.0597)	0.1462 (0.0475)	0.1427 (0.0418)	0.1821 (0.0662)	0.1425 (0.0467)
Sim3	0.1092 (0.0222)	0.0799 (0.0257)	0.0936 (0.0195)	0.0709 (0.0128)	0.0936 (0.0257)	0.0723 (0.0112)
Sim4	0.3555 (0.0702)	0.2826 (0.0691)	0.3442 (0.0804)	0.2759 (0.0639)	0.3589 (0.0638)	0.3019 (0.0779)

- The average misclassification rate increases as the missing rate rises. As expected, overlapping components typically have higher misclassification rates than the well-separated components. In addition, the fit of the parsimonious MGHD and MST models to each simulated dataset does not considerably decrease as the missing data rate rises.
- Our proposed parsimonious MGHD and MST models for incomplete data perform significantly better than their counterparts parsimonious MGHD and MST model with mean imputation method (MI/MGHD, MI/MST). In addition, our proposed parsimonious MGHD generally yields much lower misclassification rates than its competitor parsimonious MST for incomplete data when the datasets are generated from generalized hyperbolic distribution, and lower or closer misclassification rates when the datasets are generated from the skew-t or Gaussian distribution.

- Our proposed parsimonious MGHD for incomplete data generally yields similar misclassification rates under circumstances of both known clusters and unknown clusters, while its competitor parsimonious MST generally yields poorer misclassification rates with unknown clusters. This is because the BIC always finds the true number of clusters when using the MGHD for incomplete data, but tends to overestimate the number of clusters when using the MST for incomplete data for datasets with overlapping mixtures.

5.2 Italian Wine Data

In this first experiment, we apply our proposed parsimonious MGHD and MST models to the well-known Italian wine dataset, which includes thirteen chemical attributes of $n = 178$ Italian wines from Barolo (59), Grignolino (71), and Barbera (48) grape cultivars, which are treated as three intrinsic clusters. This dataset is available in the `gclus` package (Hurley, 2004) for R. This dataset is complete, so for illustration purposes we consider various levels of missing data ranging from 5 to 30% by deleting observations through an MAR mechanism. The dataset is scaled prior to analysis. The number of components is fixed at $G = 3$, then data are analyzed using our proposed parsimonious MGHD and MST models for incomplete data and their counterparts with mean imputation. The results of this analysis (Table 4) show that the parsimonious MGHD outperforms the other models for all levels of missing data.

Table 4: Misclassification rate and ARI values for our proposed approaches and using mean imputation for clustering on the wine dataset with different levels of missing rates.

r	MGHD		MST		MI/MGHD		MI/MST	
	δ	ARI	δ	ARI	δ	ARI	δ	ARI
5%	0.0506	0.8465	0.0730	0.7844	0.0562	0.8222	0.0618	0.0618
10%	0.1180	0.6779	0.1517	0.6052	0.1292	0.6455	0.1573	0.5929
20%	0.3539	0.4128	0.3427	0.4645	0.3989	0.3456	0.3764	0.3367
30%	0.3596	0.4280	0.3620	0.4073	0.3820	0.3327	0.3820	0.3327

5.3 Pima Indians Diabetes Data

Data on the diabetes status of 768 patients is obtained from the UCI Machine Learning data repository. The data include information on eight attributes, in which the attribute

of number of times pregnant is treated as continuous variable because its range is from 0 to 14. These data are a popular benchmark dataset for clustering for truly missing values, as 376 of the observations have at least one attribute missing. The data are overlapping and the numerous missing observations make clustering difficult. The detailed description of the attributes and their associated missing rates are summarized in Table 5. The dataset features 268 patients with a diabetes diagnosis and 500 without, and these are treated as two clusters. Again, this dataset is scaled prior to the analysis.

Table 5: A description of Pima Indian diabetes dataset

	No. missing values	Sample mean	Sample std. dev.
Number of times pregnant	0	3.85	3.37
Plasma glucose concentration	5	120.89	31.97
Diastolic blood pressure (mm Hg)	35	69.11	19.36
Triceps skin fold thickness (mm)	227	20.54	15.95
2-hour serum insulin(μ U/mL)	374	79.80	115.24
Body mass index	11	31.99	7.88
Diabetes pedigree function	0	0.47	0.33
Age (years)	0	33.24	11.76

Because there are two known clusters, we fix $G = 2$ and compare the BIC and ICL values for 14 covariance structures of our proposed parsimonious MGHD and MST models. The clustering results are summarized in Table 6. Lin (2014) perform a comparable cluster analysis on these via a t mixture model and matches the true cluster labels with 66.7% accuracy. Compared to Lin (2014), our proposed parsimonious MGHD model for incomplete data gives a higher accuracy rate (69.11%).

Table 6: Misclassification rate and ARI values for our proposed approaches for clustering on the Pima Indian diabetes dataset.

	Structure	BIC	ICL	δ	Accuracy
MGHD	EEE	-14016.95	-14053.61	0.3089	69.11%
MST	VVI	-14109.1	-14186.1	0.3763	62.37%

6 Discussion

Approaches for clustering incomplete data where clusters may be heavy tailed and/or asymmetric is introduced, based on MGHD and MST. These approaches were further extended to parsimonious families of MGHD and MST models via eigen-decomposition of the component scale matrices. The BIC and ICL were used for model selection. It is well known that the BIC can tend to overestimate the number of clusters in practice; however, the results presented herein show that this overestimation can sometimes be mitigated via a more flexible component density such as the MGHD. An EM algorithm was developed to fit the MGHD and MST models to incomplete data, and later implemented in R. It is worth mentioning that our approaches are also applicable in situations with no missing data; and so we have MGHD and MST analogues of the models of Celeux and Govaert (1995). Our MGHD and MST models were applied to real and simulated heterogeneous datasets for clustering in the presence of missing values, and the PMGHD family performed favourably when compared to the PMST family as well as the MGHD and MST approaches with mean imputation.

Going forward, the PMGHD and PMST approaches for clustering with missing values can easily be extended to model-based classification, discriminant analysis, and density estimation. Furthermore, Bayesian analysis via a Gibbs sampler is another popular approach to handle missing data in multivariate datasets (e.g., Lin et al., 2009), so a fully Bayesian treatment will be considered as an alternative to the EM algorithm for parameter estimation. Finally, it will also be interesting to generalize all existing approaches to developing mixture of generalized hyperbolic factor analyzer models and mixtures of multiple scaled generalized hyperbolic distributions for incomplete data (Tortora et al., 2015).

Acknowledgements This work was supported by an Ontario Graduate Scholarship (Wei), an Early Researcher Award from the Government of Ontario (McNicholas), and the Canada Research Chairs program (McNicholas).

References

Aitken, A. C. (1926). On Bernoulli’s numerical solution of algebraic equations. *Proceedings of the Royal Society of Edinburgh* 46, 289–305.

- Andrews, J. L. and P. D. McNicholas (2011). Extending mixtures of multivariate t-factor analyzers. *Statistics and Computing* 21(3), 361–373.
- Andrews, J. L. and P. D. McNicholas (2012). Model-based clustering, classification, and discriminant analysis via mixtures of multivariate t-distributions. *Statistics and Computing* 22(5), 1021–1029.
- Banfield, J. D. and A. E. Raftery (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics* 49(3), 803–821.
- Barndorff-Nielsen, O. (1977). Exponentially decreasing distributions for the logarithm of particle size. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences* 353(1674), 401–419.
- Barndorff-Nielsen, O. (1978). Hyperbolic distributions and distributions on hyperbolae. *Scandinavian Journal of Statistics* 5(3), 151–157.
- Barndorff-Nielsen, O. and P. Blæsild (1981). Hyperbolic distributions and ramifications: Contributions to theory and application. In C. Taillie, G. Patil, and B. Baldessari (Eds.), *Statistical Distributions in Scientific Work*, Volume 79 of *NATO Advanced Study Institutes Series*, pp. 19–44. Springer Netherlands.
- Barndorff-Nielsen, O. and C. Halgreen (1977). Infinite divisibility of the hyperbolic and generalized inverse Gaussian distributions. *Probability Theory and Related Fields* 38(4), 309–311.
- Biernacki, C., G. Celeux, and G. Govaert (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(7), 719–725.
- Blæsild, P. (1978). *The shape of the generalized inverse Gaussian and hyperbolic distributions*. Department of Theoretical Statistics, Institute of Mathematics, University of Aarhus.
- Böhning, D., E. Dietz, R. Schaub, P. Schlattmann, and B. Lindsay (1994). The distribution of the likelihood ratio for mixtures of densities from the one-parameter exponential family. *Annals of the Institute of Statistical Mathematics* 46(2), 373–388.
- Bouveyron, C., S. Girard, and C. Schmid (2007). High-dimensional data clustering. *Computational Statistics & Data Analysis* 52(1), 502–519.

- Branco, M. D. and D. K. Dey (2001). A general class of multivariate skew-elliptical distributions. *Journal of Multivariate Analysis* 79(1), 99 – 113.
- Browne, R. P. and P. D. McNicholas (2015). A mixture of generalized hyperbolic distributions. *Canadian Journal of Statistics* 43(2), 176–198.
- Browne, R. P., P. D. McNicholas, and C. J. Findlay (2013). A partial EM algorithm for clustering white breads. *arXiv:1302.6625*.
- Celeux, G. and G. Govaert (1995). Gaussian parsimonious clustering models. *Pattern Recognition* 28(5), 781–793.
- Dang, U. J., R. P. Browne, and P. D. McNicholas (2015). Mixtures of multivariate power exponential distributions. *Biometrics* 71(4), 1081–1089.
- Dempster, A. P., N. M. Laird, and D. B. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1–38.
- Fraley, C. and A. E. Raftery (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal* 41(8), 578–588.
- Franczak, B. C., R. P. Browne, and P. D. McNicholas (2014). Mixtures of shifted asymmetric Laplace distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 36(6), 1149–1157.
- Ghahramani, Z. and M. I. Jordan (1994). Supervised learning from incomplete data via an EM approach. In *Advances in Neural Information Processing Systems*. Citeseer.
- Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika* 40(3-4), 237–264.
- Halgreen, C. (1979). Self-decomposability of the generalized inverse Gaussian and hyperbolic distributions. *Probability Theory and Related Fields* 47(1), 13–17.
- Hubert, L. and P. Arabie (1985). Comparing partitions. *Journal of Classification* 2(1), 193–218.
- Hurley, C. B. (2004). Clustering visualizations of multidimensional data. *Journal of Computational and Graphical Statistics* 13(4), 788–806.
- Jørgensen, B. (1982). *Statistical properties of the generalized inverse Gaussian distribution*. Lecture Notes in Statistics. New York: Springer.

- Karlis, D. and A. Santourian (2009). Model-based clustering with non-elliptically contoured distributions. *Statistics and Computing* 19(1), 73–83.
- Lee, S. and G. J. McLachlan (2014). Finite mixtures of multivariate skew t-distributions: some recent and new results. *Statistics and Computing* 24(2), 181–202.
- Lin, T.-I. (2010). Robust mixture modeling using multivariate skew t distributions. *Statistics and Computing* 20(3), 343–356.
- Lin, T.-I. (2014). Learning from incomplete data via parameterized t mixture models through eigenvalue decomposition. *Computational Statistics & Data Analysis* 71, 183–195.
- Lin, T. I., H. J. Ho, and C. L. Chen (2009). Analysis of multivariate skew normal models with incomplete data. *Journal of Multivariate Analysis* 100(10), 2337–2351.
- Lin, T.-I., H. J. Ho, and P. S. Shen (2009). Computationally efficient learning of multivariate t mixture models with missing information. *Computational Statistics* 24(3), 375–392.
- Lin, T. I., J. C. Lee, and H. J. Ho (2006). On fast supervised learning for normal mixture models with missing information. *Pattern Recognition* 39(6), 1177–1187.
- Lin, T.-I. and T.-C. Lin (2011). Robust statistical modelling using the multivariate skew t distribution with complete and incomplete data. *Statistical Modelling* 11(3), 253–277.
- Lin, T.-I., P. D. McNicholas, and H. J. Ho (2014). Capturing patterns via parsimonious t mixture models. *Statistics & Probability Letters* 88, 80–87.
- Lin, T.-I., P. H. Wu, G. J. McLachlan, and S. X. Lee (2014). The skew-t factor analysis model. *arXiv 1310.5336 [stat.ME]*.
- Lindsay, B. G. (1995). Mixture Models: Theory, Geometry and Applications. In *NSF-CBMS Regional Conference Series in Probability and Statistics*, Volume 5. California: Institute of Mathematical Statistics: Hayward.
- Little, R. J. and D. B. Rubin (1987). *Statistical Analysis with Missing Data*. Wiley, New York.
- Liu, C., D. B. Rubin, and Y. N. Wu (1998). Parameter expansion to accelerate EM: The PX-EM algorithm. *Biometrika* 85(4), 755–770.
- McLachlan, G. J., D. Peel, and R. Bean (2003). Modelling high-dimensional data by mixtures of factor analyzers. *Computational Statistics and Data Analysis* 41(3), 379–388.

- McNeil, A., R. Frey, and P. Embrechts (2005). *Quantitative risk management: Concepts, Techniques and Tools*. Princeton University Press, Princeton, NJ.
- McNicholas, P. D. (2016a). *Mixture Model-Based Classification*. Boca Raton: Chapman & Hall/CRC Press.
- McNicholas, P. D. (2016b). Model-based clustering. *Journal of Classification* 33(3), 331–373.
- McNicholas, P. D. and T. B. Murphy (2008). Parsimonious Gaussian mixture models. *Statistics and Computing* 18(3), 285–296.
- McNicholas, P. D. and T. B. Murphy (2010). Model-based clustering of microarray expression data via latent Gaussian mixture models. *Bioinformatics* 26(21), 2705–2712.
- McNicholas, P. D., T. B. Murphy, A. F. McDaid and D. Frost (2010). Serial and parallel implementations of model-based clustering via parsimonious Gaussian mixture models. *Computational Statistics and Data Analysis* 54(3), 711–723.
- Morris, K. and P. D. McNicholas (2016). Clustering, classification, discriminant analysis, and dimension reduction via generalized hyperbolic mixtures. *Computational Statistics & Data Analysis* 97, 133–150.
- Murray, P. M., R. P. Browne, and P. D. McNicholas (2014a). Mixtures of skew-factor analyzers. *Computational Statistics & Data Analysis* 77, 326–335.
- Murray, P. M., P. D. McNicholas, and R. P. Browne (2014b). A mixture of common skew-t factor analysers. *Stat* 3(1), 68–82.
- O’Hagan, A., T. B. Murphy, I. C. Gormley, P. D. McNicholas, and D. Karlis (2016). Clustering with the multivariate normal inverse Gaussian distribution. *Computational Statistics & Data Analysis* 93, 18–30.
- Peel, D. and G. J. McLachlan (2000). Robust mixture modelling using the t distribution. *Statistics and Computing* 10(4), 339–348.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika* 63(3), 581–592.
- Sahu, S. K., D. K. Dey, and M. D. Branco (2003). A new class of multivariate skew distributions with applications to Bayesian regression models. *Canadian Journal of Statistics* 31(2), 129–150.
- Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* 6, 461–464.

- Tiedeman, D. V. (1955). On the study of types. In S. B. Sells (Ed.), *Symposium on Pattern Analysis*. Randolph Field, Texas: Air University, U.S.A.F. School of Aviation Medicine.
- Tortora, C., P. D. McNicholas, and R. P. Browne (2015). A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification*, 1–18.
- Tortora, C., P. D. McNicholas, and R. P. Browne (2016). A mixture of generalized hyperbolic factor analyzers. *Advances in Data Analysis and Classification* 10(4), 423–440.
- Vrbik, I. and P. McNicholas (2012). Analytic calculations for the EM algorithm for multivariate skew-t mixture models. *Statistics & Probability Letters* 82(6), 1169–1174.
- Wang, H. X., Q. B. Zhang, B. Luo, and S. Wei (2004). Robust mixture modelling using multivariate t-distribution with missing information. *Pattern Recognition Letters* 25(6), 701–710.

A Some Matrix Computations

We here present some useful matrix computation results that are employed in the derivation of the conditional pdf of a partitioned generalized hyperbolic and multivariate skew-t random vector \mathbf{X} in Propositions 3 and 6.

Consider a partitioned random vector \mathbf{X} of p -dimension that follows the pdf as in Equation (9) with

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_1 \\ \mathbf{X}_2 \end{pmatrix} \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix}, \quad (22)$$

where \mathbf{X}_1 and \mathbf{X}_2 have dimensions d_1 and $d_2 = p - d_1$, respectively. The mean, skewness and dispersion matrix are composed of blocks of appropriate dimensions as partitions of \mathbf{X} . Sometimes, it is more convenient to work with the inverse of dispersion matrix $\boldsymbol{\Sigma}^{-1}$:

$$\boldsymbol{\Sigma}^{-1} = \begin{pmatrix} (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12}\boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}^T)^{-1} & -\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \\ -(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1}\boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1} & (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12})^{-1} \end{pmatrix}. \quad (23)$$

Furthermore, we have for the determinant of $\boldsymbol{\Sigma}$:

$$\det(\boldsymbol{\Sigma}) = \det(\boldsymbol{\Sigma}_{11})\det(\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^T\boldsymbol{\Sigma}_{11}^{-1}\boldsymbol{\Sigma}_{12}). \quad (24)$$

B Outline of Proof of Proposition 3

Here, we derive the conditional density of \mathbf{X}_2 given that $\mathbf{X}_1 = \mathbf{x}_1$ if \mathbf{X}_1 and \mathbf{X}_2 are jointly generalized hyperbolic distributed, i.e., $\mathbf{X} \sim \text{GHD}_p(\lambda, \omega, \boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\beta})$ with the partition in Appendix A. Although basic probability theory indicates that the conditional pdf is a ratio of the joint and marginal pdfs, the expression takes a very complicated form. The results from Appendix A are heavily used in the course of the derivations. The conditional density is given by

$$\begin{aligned} f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) &= \frac{f_{\mathbf{X}_1, \mathbf{X}_2}(\mathbf{x}_1, \mathbf{x}_2)}{f_{\mathbf{X}_1}(\mathbf{x}_1)} \\ &= \frac{\left[\frac{\omega + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma})}{\omega + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}} \right]^{\frac{\lambda - p/2}{2}} K_{\lambda - p/2} \left(\sqrt{(\omega + \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}))(\omega + \boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta})} \right)}{\left[\frac{\omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11})}{\omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1} \right]^{\frac{\lambda - d_1/2}{2}} K_{\lambda - d_1/2} \left(\sqrt{(\omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11}))(\omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1)} \right)} \frac{(2\pi)^{p/2} |\boldsymbol{\Sigma}|^{1/2} K_\lambda(\omega) \exp\{-(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}\}}{(2\pi)^{d_1/2} |\boldsymbol{\Sigma}_{11}|^{1/2} K_\lambda(\omega) \exp\{-(\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1\}}, \end{aligned}$$

where we combine (9) and Proposition 2. For the moment, we focus on the linear form and quadratic form in which \mathbf{x} enters the pdf in (9). Inserting the partition of \mathbf{X} , $\boldsymbol{\mu}$, $\boldsymbol{\beta}$, and $\boldsymbol{\Sigma}$ in (22) and the inverse of dispersion matrix $\boldsymbol{\Sigma}^{-1}$ (23) into the quadratic form yields

$$\begin{aligned} \delta(\mathbf{x}, \boldsymbol{\mu} | \boldsymbol{\Sigma}) &= (\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \\ &= \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \mathbf{x}_1 - \boldsymbol{\mu}_1 \\ \mathbf{x}_2 - \boldsymbol{\mu}_2 \end{pmatrix} \\ &= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top (\boldsymbol{\Sigma}_{11} - \boldsymbol{\Sigma}_{12} \boldsymbol{\Sigma}_{22}^{-1} \boldsymbol{\Sigma}_{12}^\top)^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &\quad - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &\quad - (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &\quad + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &\quad - (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &\quad - (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12} (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2) \\ &= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1) \\ &\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1))^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\mathbf{x}_2 - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1)) \\ &= \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11}) + \delta(\mathbf{x}_2, \boldsymbol{\mu}_{2|1} | \boldsymbol{\Sigma}_{2|1}), \end{aligned} \tag{25}$$

where $\boldsymbol{\mu}_{2|1} = \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1)$ and $\boldsymbol{\Sigma}_{2|1} = (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1}$.

Similarly, inserting into the linear form, following the same algebra as above, yields

$$\begin{aligned}
(\mathbf{x} - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} &= \begin{pmatrix} (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top & (\mathbf{x}_2 - \boldsymbol{\mu}_2)^\top \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1 \\
&\quad + (\mathbf{x}_2 - \boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1}(\mathbf{x}_1 - \boldsymbol{\mu}_1))^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1) \\
&= (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1 + (\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})^\top \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\beta}_{2|1},
\end{aligned} \tag{26}$$

where $\boldsymbol{\mu}_{2|1}$ and $\boldsymbol{\Sigma}_{2|1}$ are as described above, and $\boldsymbol{\beta}_{2|1} = \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1$.

Furthermore, we investigate the term $\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}$, we obtain

$$\begin{aligned}
\boldsymbol{\beta}^\top \boldsymbol{\Sigma}^{-1} \boldsymbol{\beta} &= \begin{pmatrix} \boldsymbol{\beta}_1^\top & \boldsymbol{\beta}_2^\top \end{pmatrix} \boldsymbol{\Sigma}^{-1} \begin{pmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{pmatrix} \\
&= \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1 + (\boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1)^\top (\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12})^{-1} (\boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1) \\
&= \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2|1}^\top \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\beta}_{2|1}.
\end{aligned} \tag{27}$$

Finally, we substitute (24), (25), (26), and (27), and $p = d_1 + d_2$ into the conditional density, and after some simple linear algebra, we obtain

$$\begin{aligned}
f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) &= \frac{\left(\frac{\omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11}) + \delta(\mathbf{x}_2, \boldsymbol{\mu}_{2|1} | \boldsymbol{\Sigma}_{2|1})}{\omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2|1}^\top \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\beta}_{2|1}} \right)^{\frac{\lambda - \frac{d_1}{2} - \frac{d_2}{2}}{2}} \left[\frac{\omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1}{\omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11})} \right]^{\frac{\lambda - d_1/2}{2}}}{(2\pi)^{\frac{d_2}{2}} |\boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}|^{\frac{1}{2}}} \times \\
&\frac{K_{\lambda - \frac{d_1}{2} - \frac{d_2}{2}} \left(\sqrt{(\omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11})) + \delta(\mathbf{x}_2, \boldsymbol{\mu}_{2|1} | \boldsymbol{\Sigma}_{2|1})} (\omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1 + \boldsymbol{\beta}_{2|1}^\top \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\beta}_{2|1}) \right)}{K_{\lambda - \frac{d_1}{2}} \left(\sqrt{(\omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11}))} (\omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1) \right) \exp(-(\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})^\top \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\beta}_{2|1})}.
\end{aligned}$$

Set $\lambda_{2|1} = \lambda - \frac{d_1}{2}$, $\chi_{2|1} = \omega + \delta(\mathbf{x}_1, \boldsymbol{\mu}_1 | \boldsymbol{\Sigma}_{11})$, and $\psi_{2|1} = \omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1$, then we obtain

$$f_{\mathbf{X}_2|\mathbf{X}_1}(\mathbf{x}_2 | \mathbf{x}_1) = \left[\frac{\chi_{2|1} + \delta(\mathbf{x}_2, \boldsymbol{\mu}_{2|1} | \boldsymbol{\Sigma}_{2|1})}{\psi_{2|1} + \boldsymbol{\beta}_{2|1}^\top \boldsymbol{\Sigma}_{2|1} \boldsymbol{\beta}_{2|1}} \right]^{\frac{\lambda_{2|1} - \frac{d_2}{2}}{2}} \times \\ \frac{\left(\frac{\psi_{2|1}}{\chi_{2|1}} \right)^{\frac{\lambda_{2|1}}{2}} K_{\lambda_{2|1} - \frac{d_2}{2}} \left(\sqrt{(\psi_{2|1} + \boldsymbol{\beta}_{2|1}^\top \boldsymbol{\Sigma}_{2|1} \boldsymbol{\beta}_{2|1})(\chi_{2|1} + \delta(\mathbf{x}_2, \boldsymbol{\mu}_{2|1} | \boldsymbol{\Sigma}_{2|1}))} \right)}{(2\pi)^{\frac{d_2}{2}} |\boldsymbol{\Sigma}_{2|1}|^{\frac{1}{2}} K_{\lambda_{2|1}}(\sqrt{\chi_{2|1} \psi_{2|1}}) \exp(-(\mathbf{x}_2 - \boldsymbol{\mu}_{2|1})^\top \boldsymbol{\Sigma}_{2|1}^{-1} \boldsymbol{\beta}_{2|1})}.$$

Comparison with (6) reveals that this is a generalized hyperbolic distribution in the parameterization of McNeil et al. (2005) with

$$\begin{aligned} \lambda_{2|1} &= \lambda - \frac{d_1}{2}, & \chi_{2|1} &= \omega + (\mathbf{x}_1 - \boldsymbol{\mu}_1)^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ \psi_{2|1} &= \omega + \boldsymbol{\beta}_1^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1, & \boldsymbol{\mu}_{2|1} &= \boldsymbol{\mu}_2 + \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} (\mathbf{x}_1 - \boldsymbol{\mu}_1), \\ \boldsymbol{\Sigma}_{2|1} &= \boldsymbol{\Sigma}_{22} - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\Sigma}_{12}, & \boldsymbol{\beta}_{2|1} &= \boldsymbol{\beta}_2 - \boldsymbol{\Sigma}_{12}^\top \boldsymbol{\Sigma}_{11}^{-1} \boldsymbol{\beta}_1. \end{aligned}$$

C MST with Incomplete Data

Analogous to the MGHD model (14), the MST model takes the density

$$f_{\text{MST}}(\mathbf{X}_i | \boldsymbol{\Theta}) = \sum_{g=1}^G \pi_g f_{\text{ST}}(\mathbf{X}_i | v_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g), \quad (28)$$

where $\boldsymbol{\Theta} = (\pi, \mathbf{v}_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g, \boldsymbol{\beta}_g)$ with $\mathbf{v}_g = (v_1, \dots, v_g)$ and $\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$, and $\boldsymbol{\beta}_g$ are as defined above. By introducing the group membership variables $\mathbf{Z}_i \sim \mathcal{M}(1; \pi_1, \dots, \pi_G)$, convenient three-layer hierarchical representations are given by

$$\begin{aligned} \mathbf{X}_i | (w_{ig}, Z_{ig} = 1) &\sim \mathcal{N}(\boldsymbol{\mu}_g + w_{ig} \boldsymbol{\beta}_g, w_{ig} \boldsymbol{\Sigma}_g) \\ W_{ig} | (Z_{ig} = 1) &\sim \text{IG}(v_g/2, v_g/2). \\ \mathbf{Z}_i &\sim \mathcal{M}(1; \pi_1, \dots, \pi_G) \end{aligned} \quad (29)$$

Assume that the matrix $\mathbf{X} = (\mathbf{X}^{\text{ot}}, \mathbf{X}^{\text{mt}})^\top$ contains missing data. For each $\mathbf{x}_i = (\mathbf{x}_i^{\text{ot}}, \mathbf{x}_i^{\text{mt}})^\top$, we write $\boldsymbol{\mu}_g = (\boldsymbol{\mu}_{g,i}^{\text{ot}}, \boldsymbol{\mu}_{g,i}^{\text{mt}})^\top$, $\boldsymbol{\beta}_g = (\boldsymbol{\beta}_{g,i}^{\text{ot}}, \boldsymbol{\beta}_{g,i}^{\text{mt}})^\top$, and finally the g th dispersion matrix $\boldsymbol{\Sigma}_g$ is partitioned as in (17). Hence, based on (29), we have the following conditional distributions:

- The marginal distribution of \mathbf{X}_i^o is

$$\mathbf{X}_i^o \sim \sum_{g=1}^G \pi_g f_{\text{ST}, p_i^o}(\lambda_g, \omega_g, \boldsymbol{\mu}_{g,i}^o, \boldsymbol{\Sigma}_{g,i}^{oo}, \boldsymbol{\beta}_{g,i}^o),$$

where p_i^o is the dimension corresponding to the observed component \mathbf{x}_i^o , which should be exactly written as $p_i^{o_i}$ but here is simplified.

- The conditional distribution of \mathbf{X}_i^m given \mathbf{x}_i^o and $Z_{ig} = 1$, according to Proposition 6, is

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, Z_{ig} = 1 \sim \text{GH}_{p-p_i^o}(\lambda_{g,i}^{m|o}, \chi_{g,i}^{m|o}, \psi_{g,i}^{m|o}, \boldsymbol{\mu}_{g,i}^{m|o}, \boldsymbol{\Sigma}_{g,i}^{m|o}, \boldsymbol{\beta}_{g,i}^{m|o}), \quad (30)$$

where

$$\begin{aligned} \lambda_{g,i}^{m|o} &= -\frac{v_g + p_i^o}{2}, & \psi_{g,i}^{m|o} &= v_g + (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o), \\ \psi_{g,i}^{m|o} &= \boldsymbol{\beta}_{g,i}^{o\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o, & \boldsymbol{\mu}_{g,i}^{m|o} &= \boldsymbol{\mu}_{g,i}^m + \boldsymbol{\Sigma}_{g,i}^{om\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o), \\ \boldsymbol{\Sigma}_{g,i}^{m|o} &= \boldsymbol{\Sigma}_{g,i}^{mm} - \boldsymbol{\Sigma}_{g,i}^{om\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\Sigma}_{g,i}^{om}, & \boldsymbol{\beta}_{g,i}^{m|o} &= \boldsymbol{\beta}_{g,i}^m - \boldsymbol{\Sigma}_{g,i}^{om\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o. \end{aligned}$$

- The conditional distribution of \mathbf{X}_i^m given \mathbf{x}_i^o, w_{ig} , and $Z_{ig} = 1$ is

$$\mathbf{X}_i^m \mid \mathbf{x}_i^o, w_{ig}, Z_{ig} = 1 \sim \mathcal{N}_{p-p_i^o}(\boldsymbol{\mu}_{g,i}^{m|o} + w_{ig} \boldsymbol{\beta}_{g,i}^{m|o}, w_{ig} \boldsymbol{\Sigma}_{g,i}^{m|o}). \quad (31)$$

- The conditional distribution of W_i given \mathbf{x}_i^o and $Z_{ig} = 1$ is

$$W_{ig} \mid \mathbf{x}_i^o, Z_{ig} = 1 \sim \text{GIG} \left(\boldsymbol{\beta}_{g,i}^{o\top} (\boldsymbol{\Sigma}_{g,i}^{oo})^{-1} \boldsymbol{\beta}_{g,i}^o, v_g + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^o \mid \boldsymbol{\Sigma}_{g,i}^{oo}), -\frac{v_g + p_i^o}{2} \right). \quad (32)$$

As in the case of the MGHD model with incomplete data, the complete data consists of the observed \mathbf{x}_i , the missing group membership z_{ig} , the latent w_{ig} , as well as the actual missing data \mathbf{x}_i^m , for $i = 1, \dots, n$ and $g = 1, \dots, G$. Again, the complete data log-likelihood function is given by

$$l_c(\boldsymbol{\Theta}) = \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[\log \pi_g + \log \phi(\mathbf{x}_i^o, \mathbf{x}_i^m \mid \boldsymbol{\mu}_g + w_{ig} \boldsymbol{\beta}_g, w_{ig} \boldsymbol{\Sigma}_g) + \log f_{\text{IG}}(w_{ig} \mid v_g/2, v_g/2) \right]. \quad (33)$$

Furthermore, one can simplify (33) to

$$\begin{aligned}
l_c(\Theta) = & \sum_{i=1}^n \sum_{g=1}^G z_{ig} \log \pi_g + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[-\frac{p}{2} \log(2\pi) - \frac{p}{2} \log w_{ig} + \frac{1}{2} \log |\Sigma_g^{-1}| \right] \\
& - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\Sigma_g^{-1} z_{ig} \frac{1}{w_{ig}} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \\ (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o) & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)(\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\
& + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\Sigma_g^{-1} z_{ig} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^o \\ \boldsymbol{\beta}_{g,i}^m \end{pmatrix} \begin{pmatrix} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o)^\top & (\mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m)^\top \end{pmatrix} \right) \\
& + \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G \text{tr} \left(\Sigma_g^{-1} z_{ig} \begin{pmatrix} \mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^o \\ \mathbf{x}_i^m - \boldsymbol{\mu}_{g,i}^m \end{pmatrix} \begin{pmatrix} \boldsymbol{\beta}_{g,i}^{o\top} & \boldsymbol{\beta}_{g,i}^{m\top} \end{pmatrix} \right) - \frac{1}{2} \sum_{i=1}^n \sum_{g=1}^G z_{ig} w_{ig} \boldsymbol{\beta}_{g,i}^\top \Sigma_g^{-1} \boldsymbol{\beta}_{g,i} \\
& + \sum_{i=1}^n \sum_{g=1}^G z_{ig} \left[\frac{v_g}{2} \log \left(\frac{v_g}{2} \right) - \log \Gamma \left(\frac{v_g}{2} \right) - \left(\frac{v_g}{2} + 1 \right) \log w_{ig} - \frac{v_g}{2w_{ig}} \right]
\end{aligned} \tag{34}$$

On the k th iteration of the E-step, the expected value of the complete-data log-likelihood is computed given the observed data \mathbf{X}^o and the current parameter updates $\Theta^{(k)}$. Denote by $\tau_{ig}^{(k)}$ the *a posteriori* probability that the i th observation belongs to the g th component of the mixture. Specifically, it can be calculated as

$$\tau_{ig}^{(k+1)} := \mathbb{E}(Z_{ig} \mid \mathbf{x}_i^o, \Theta^{(k)}) = \frac{\pi_g^{(k)} f_{\text{ST}, p_i^o}(\mathbf{x}_i^o; v_g^{(k)}, \boldsymbol{\mu}_{g,i}^{o(k)}, \boldsymbol{\Sigma}_{g,i}^{oo(k)}, \boldsymbol{\beta}_{g,i}^{o(k)})}{\sum_{l=1}^G \pi_l^{(k)} f_{\text{ST}, p_i^o}(\mathbf{x}_i^o; v_l^{(k)}, \boldsymbol{\mu}_{l,i}^{o(k)}, \boldsymbol{\Sigma}_{l,i}^{oo(k)}, \boldsymbol{\beta}_{l,i}^{o(k)})}.$$

Given the observed data \mathbf{x}^o , the current parameter updates $\Theta^{(k)}$, and conditional distributions (30) and (32), taking expectations for (34) leads to the following expectation updates in the E-step:

$$\begin{aligned}
A_{ig}^{(k)} &:= \mathbb{E}(W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \sqrt{\frac{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}}} \\
&\quad \times \frac{K_{-(v_g^{(k)} + p_i^o)/2+1} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{-(v_g^{(k)} + p_i^o)/2} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}}, \\
B_{ig}^{(k)} &:= \mathbb{E}(1/W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) \\
&= \frac{v_g^{(k)} + p_i^o}{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})} + \sqrt{\frac{\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}} \\
&\quad \times \frac{K_{-(v_g^{(k)} + p_i^o)/2+1} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}{K_{-(v_g^{(k)} + p_i^o)/2} \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right)}, \\
C_{ig}^{(k)} &:= \mathbb{E}(\log W_{ig} \mid \mathbf{x}_i^o, z_{ig} = 1; \boldsymbol{\Theta}^{(k)}) = \log \left(\sqrt{\frac{v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})}{\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)}}}} \right) \\
&\quad + \frac{\partial}{\partial t} \log \left\{ K_t \left(\sqrt{(v_g^{(k)} + \delta(\mathbf{x}_i^o, \boldsymbol{\mu}_{g,i}^{o(k)} \mid \boldsymbol{\Sigma}_{g,i}^{oo(k)})) (\boldsymbol{\beta}_{g,i}^{o(k)\top} (\boldsymbol{\Sigma}_{g,i}^{oo(k)})^{-1} \boldsymbol{\beta}_{g,i}^{o(k)})} \right) \right\} \Big|_{t=-(v_g^{(k)} + p_i^o)/2}, \\
\hat{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}(\mathbf{X}_i^m \mid \mathbf{x}_i^o, Z_{ig} = 1) = \boldsymbol{\mu}_{g,i}^{m|o(k)} + A_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)}, \\
\tilde{\mathbf{x}}_{ig}^{m(k)} &:= \mathbb{E}((1/W_i) \mathbf{X}_i^m \mid \mathbf{x}_i^o, Z_{ig} = 1) = B_{ig}^{(k)} \boldsymbol{\mu}_{g,i}^{m|o(k)} + \boldsymbol{\beta}_{g,i}^{m|o(k)}, \\
\tilde{\tilde{\mathbf{x}}}_{ig}^{m(k)} &:= \mathbb{E}((1/w_i) \mathbf{X}_i^m \mathbf{X}_i^{m\top} \mid \mathbf{x}_i^o, Z_{ig} = 1) = \boldsymbol{\Sigma}_{g,i}^{m|o(k)} + B_{ig}^{(k)} \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top \\
&\quad + \boldsymbol{\mu}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top + \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\mu}_{g,i}^{m|o(k)})^\top + A_{ig}^{(k)} \boldsymbol{\beta}_{g,i}^{m|o(k)} (\boldsymbol{\beta}_{g,i}^{m|o(k)})^\top.
\end{aligned}$$

For convenience, let $n_g^{(k)} = \sum_{i=1}^n \tau_{ig}^{(k)}$, $\bar{A}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \tau_{ig}^{(k)} A_{ig}^{(k)}$, $\bar{B}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \tau_{ig}^{(k)} B_{ig}^{(k)}$, and $\bar{C}_g^{(k)} = 1/n_g^{(k)} \sum_{i=1}^n \tau_{ig}^{(k)} C_{ig}^{(k)}$. On the k th iteration of the M-step, we get updates for the parameter estimates of the mixture as follows:

$$\begin{aligned}
\pi_g^{(k+1)} &= \frac{n_g^{(k)}}{n}, \\
\boldsymbol{\mu}_g^{(k+1)} &= \frac{1}{\sum_{i=1}^n \hat{\tau}_{ig}^{(k)} (\bar{A}_g^{(k)} B_{ig}^{(k)} - 1)} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)} \left((\bar{A}_g^{(k)} B_{ig}^{(k)} - 1) \mathbf{x}_i^o \right), \\
\boldsymbol{\beta}_g^{(k+1)} &= \frac{1}{\sum_{i=1}^n \hat{\tau}_{ig}^{(k)} (\bar{A}_g^{(k)} B_{ig}^{(k)} - 1)} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)} \left((\bar{B}_g^{(k)} - B_{ig}^{(k)}) \mathbf{x}_i^o \right), \\
\boldsymbol{\Sigma}_g^{(k+1)} &= \frac{1}{n_g^{(k)}} \sum_{i=1}^n \hat{\tau}_{ig}^{(k)} \boldsymbol{\Sigma}_{ig}^{(k+1)} - (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)}) \boldsymbol{\beta}_g^{(k+1)\top} - \boldsymbol{\beta}_g^{(k+1)} (\bar{\mathbf{x}}_g - \boldsymbol{\mu}_g^{(k+1)})^\top + \bar{A}_g^{(k+1)} \boldsymbol{\beta}_g^{(k+1)} \boldsymbol{\beta}_g^{(k+1)\top},
\end{aligned}$$

where

$$\begin{aligned}
\bar{\mathbf{x}}_g &= \frac{1}{n_g^{(k+1)}} \sum_{i=1}^n \hat{\tau}_{ig}^{(k+1)} \begin{pmatrix} \mathbf{x}_i^o \\ \tilde{\mathbf{x}}_{ig}^{m(k+1)} \end{pmatrix}, \\
\boldsymbol{\Sigma}_{ig}^{(k+1)} &= \begin{pmatrix} B_{ig}^{(k+1)} (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^{o(k+1)}) (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^{o(k+1)})^\top & (\mathbf{x}_i^o - \hat{\boldsymbol{\mu}}_g^{o(k+1)}) (\tilde{\mathbf{x}}_{ig}^{m(k+1)} - B_{ig}^{(k+1)} \hat{\boldsymbol{\mu}}_g^{m(k+1)})^\top \\ (\tilde{\mathbf{x}}_{ig}^{m(k+1)} - B_{ig}^{(k+1)} \hat{\boldsymbol{\mu}}_g^{m(k+1)}) (\mathbf{x}_i^o - \boldsymbol{\mu}_{g,i}^{o(k+1)})^\top & \mathbf{k}_{ig}^{m(k+1)} \end{pmatrix},
\end{aligned}$$

where

$$\mathbf{k}_{ig}^{m(k+1)} = \tilde{\mathbf{x}}_{ig}^{m(k+1)} - \tilde{\mathbf{x}}_{ig}^{m(k)} \hat{\boldsymbol{\mu}}_g^{m(k+1)T} - \hat{\boldsymbol{\mu}}_g^{m(k+1)} \tilde{\mathbf{x}}_i^{m(k)\top} + B_{ig}^{(k)} \hat{\boldsymbol{\mu}}_g^{m(k+1)} \hat{\boldsymbol{\mu}}_g^{m(k+1)\top}.$$

Finally, as for the degree of freedom parameter v_g , the update does not exist in closed form. The update $v_g^{(k+1)}$ is the solution of

$$\log \left(\frac{v_g^{(k+1)}}{2} \right) + 1 - \varphi \left(\frac{v_g^{(k+1)}}{2} \right) - \frac{1}{n_g^{(k)}} \sum_{i=1}^n \tau_{ig} (C_{ig}^{(k)} + B_{ig}^{(k)}) = 0, \quad (35)$$

where $\varphi(\cdot)$ is the digamma function.

D Results from Simulation Studies

The results from the simulation studies are summarized in Tables 7 and 8.

Table 7: A comparison of average misclassification rates between MGHD, MST, and MI models with standard deviations in parentheses (replications=20) with $G = 2$.

	r	MGHD		MST		MI/MGHD		MI/MST	
		δ	ARI	δ	ARI	δ	ARI	δ	ARI
Sim1	5%	0.0539 (0.016)	0.7963 (0.0553)	0.1346 (0.0293)	0.5362 (0.0868)	0.111 (0.0259)	0.6039 (0.0807)	0.1309 (0.0339)	0.5482 (0.0972)
	10%	0.0503 (0.0135)	0.8093 (0.0484)	0.1344 (0.0325)	0.3355 (0.1934)	0.1684 (0.0302)	0.3355 (0.1934)	0.233 (0.1167)	0.3355 (0.1934)
	20%	0.0641 (0.0208)	0.761 (0.0701)	0.1241 (0.0409)	0.2686 (0.2561)	0.2821 (0.1452)	0.2686 (0.2561)	0.3385 (0.1135)	0.1516 (0.1546)
	30%	0.0684 (0.0289)	0.7478 (0.0933)	0.113 (0.0318)	0.602 (0.1001)	0.1035 (0.0925)	0.6606 (0.1884)	0.3338 (0.1836)	0.2374 (0.2969)
Sim2	5%	0.1095 (0.0335)	0.6133 (0.0961)	0.1676 (0.0575)	0.4532 (0.1382)	0.1998 (0.053)	0.3699 (0.1175)	0.195 (0.0635)	0.386 (0.132)
	10%	0.1114 (0.0462)	0.4481 (0.146)	0.1694 (0.0563)	0.4481 (0.146)	0.2621 (0.0678)	0.2422 (0.1088)	0.2893 (0.0966)	0.2116 (0.1632)
	20%	0.1244 (0.0274)	0.5662 (0.0786)	0.1786 (0.0546)	0.4232 (0.1225)	0.1853 (0.1014)	0.4342 (0.1698)	0.2924 (0.0993)	0.2084 (0.1561)
	30%	0.1244 (0.0297)	0.5667 (0.0874)	0.172 (0.0426)	0.436 (0.11)	0.1293 (0.0356)	0.5536 (0.0897)	0.2616 (0.1529)	0.3147 (0.2266)
Sim3	5%	0.0208 (0.0049)	0.9186 (0.0187)	0.0454 (0.0288)	0.9186 (0.0187)	0.0349 (0.0045)	0.8651 (0.0167)	0.0938 (0.0915)	0.6913 (0.1774)
	10%	0.0304 (0.0054)	0.882 (0.0204)	0.0531 (0.0286)	0.8014 (0.1006)	0.0611 (0.0103)	0.7703 (0.0363)	0.1384 (0.1163)	0.5736 (0.2459)
	20%	0.0497 (0.01)	0.8131 (0.0365)	0.0689 (0.0272)	0.7373 (0.0971)	0.1461 (0.1122)	0.5516 (0.1941)	0.3017 (0.1261)	0.2199 (0.2232)
	30%	0.0674 (0.0091)	0.6472 (0.1719)	0.1076 (0.0921)	0.7483 (0.0315)	0.2808 (0.1631)	0.292 (0.2823)	0.4618 (0.037)	0.009 (0.0169)
Sim4	5%	0.191 (0.0454)	0.3883 (0.1057)	0.2891 (0.0789)	0.1997 (0.135)	0.2065 (0.0566)	0.3553 (0.1238)	0.2968 (0.1084)	0.3553 (0.1238)
	10%	0.293 (0.1201)	0.2248 (0.19)	0.3025 (0.073)	0.1745 (0.1055)	0.2543 (0.0965)	0.2756 (0.1586)	0.3133 (0.1041)	0.1789 (0.1505)
	20%	0.272 (0.0942)	0.2403 (0.1339)	0.3004 (0.0917)	0.1896 (0.1353)	0.3101 (0.1175)	0.1953 (0.1531)	0.317 (0.1056)	0.1748 (0.1315)
	30%	0.2748 (0.0575)	0.2138 (0.1005)	0.3241 (0.0776)	0.1447 (0.0958)	0.415 (0.0939)	0.0605 (0.112)	0.4699 (0.051)	0.0114 (0.0502)
Sim5	5%	0.0776 (0.0214)	0.7146 (0.0714)	0.1155 (0.0399)	0.5965 (0.1201)	0.1448 (0.0549)	0.5151 (0.1374)	0.118 (0.0274)	0.5855 (0.0836)
	10%	0.0783 (0.0328)	0.7149 (0.1067)	0.1214 (0.0388)	0.5782 (0.1181)	0.1816 (0.0483)	0.4129 (0.1173)	0.1665 (0.0377)	0.4489 (0.0954)
	20%	0.0836 (0.0378)	0.6982 (0.1204)	0.1124 (0.0411)	0.6065 (0.1272)	0.2556 (0.169)	0.3462 (0.2948)	0.2638 (0.1011)	0.2605 (0.1796)
	30%	0.101 (0.0478)	0.6447 (0.145)	0.0986 (0.0298)	0.6469 (0.0946)	0.1441 (0.1458)	0.5864 (0.2624)	0.2673 (0.1903)	0.3536 (0.3161)
Sim6	5%	0.2235 (0.0493)	0.3136 (0.0996)	0.2199 (0.0704)	0.3312 (0.1132)	0.2749 (0.0806)	0.226 (0.1357)	0.2469 (0.075)	0.2761 (0.1141)
	10%	0.2439 (0.08)	0.2853 (0.1459)	0.2384 (0.0882)	0.3019 (0.1451)	0.2813 (0.0916)	0.2219 (0.1348)	0.2784 (0.0854)	0.2227 (0.1338)
	20%	0.2518 (0.0508)	0.2548 (0.0966)	0.3039 (0.0997)	0.19 (0.1257)	0.4419 (0.0517)	0.0216 (0.0377)	0.4409 (0.0416)	0.0182 (0.0347)
	30%	0.2495 (0.0749)	0.2709 (0.1285)	0.241 (0.0477)	0.2755 (0.0935)	0.2145 (0.0355)	0.3292 (0.0778)	0.2975 (0.0975)	0.1987 (0.1153)

Table 8: A comparison of average misclassification rates between MGH, MST, and MI models with standard deviations in parentheses (replications=20) with $G = 1, \dots, 4$.

Datasets	r	MGH		MST		MI/MGH		MI/MST	
		δ	ARI	δ	ARI	δ	ARI	δ	ARI
Sim1	5%	0.0608 (0.0292)	0.7744 (0.0925)	0.0688 (0.0557)	0.7712 (0.0998)	0.1206 (0.0302)	0.5935 (0.0874)	0.1185 (0.0341)	0.6069 (0.098)
	10%	0.0578 (0.0116)	0.7823 (0.0412)	0.2769 (0.0895)	0.4558 (0.2147)	0.1879 (0.0392)	0.5029 (0.109)	0.2325 (0.0882)	0.4794 (0.139)
	20%	0.0674 (0.0335)	0.7523 (0.1082)	0.2311 (0.0604)	0.5615 (0.1052)	0.3108 (0.0552)	0.2975 (0.1387)	0.2963 (0.0541)	0.3703 (0.1209)
	30%	0.0746 (0.0309)	0.7267 (0.099)	0.2369 (0.0576)	0.5605 (0.075)	0.4265 (0.1155)	0.2461 (0.1705)	0.3936 (0.0824)	0.2825 (0.1374)
Sim2	5%	0.1114 (0.0398)	0.6092 (0.1061)	0.3174 (0.0936)	0.3703 (0.1716)	0.1769 (0.0534)	0.4482 (0.1331)	0.3348 (0.0986)	0.3444 (0.1437)
	10%	0.1188 (0.0425)	0.5873 (0.1126)	0.4068 (0.1431)	0.4068 (0.1431)	0.3018 (0.1263)	0.3018 (0.1263)	0.3281 (0.1149)	0.3281 (0.1149)
	20%	0.1240 (0.0444)	0.5722 (0.1225)	0.3103 (0.095)	0.4056 (0.1076)	0.3153 (0.0531)	0.3081 (0.1011)	0.3354 (0.0648)	0.294 (0.1311)
	30%	0.1319 (0.0437)	0.5482 (0.1121)	0.3036 (0.0577)	0.386 (0.069)	0.476 (0.0869)	0.1319 (0.1268)	0.4586 (0.086)	0.1723 (0.1109)
Sim3	5%	0.0198 (0.0055)	0.9227 (0.0211)	0.2155 (0.0774)	0.6765 (0.1619)	0.0335 (0.0075)	0.8765 (0.0284)	0.2608 (0.108)	0.5412 (0.2765)
	10%	0.0556 (0.1049)	0.8316 (0.1979)	0.2386 (0.1161)	0.5548 (0.2822)	0.0878 (0.0367)	0.7565 (0.0503)	0.2969 (0.1256)	0.3988 (0.2891)
	20%	0.0744 (0.1010)	0.7629 (0.1855)	0.246 (0.0793)	0.3157 (0.2448)	0.3251 (0.1154)	0.3157 (0.2448)	0.2673 (0.0673)	0.4629 (0.1664)
	30%	0.0769 (0.0141)	0.7162 (0.0476)	0.2473 (0.0938)	0.0904 (0.1201)	0.4741 (0.0726)	0.0904 (0.1201)	0.5329 (0.1012)	0.09 (0.1422)
Sim4	5%	0.2419 (0.1054)	0.3074 (0.1699)	0.441 (0.0632)	0.1751 (0.1035)	0.2066 (0.055)	0.355 (0.1256)	0.3004 (0.1066)	0.1875 (0.0965)
	10%	0.3004 (0.1066)	0.2011 (0.1593)	0.4401 (0.0608)	0.1519 (0.075)	0.2518 (0.0826)	0.2938 (0.1387)	0.4048 (0.0689)	0.1842 (0.0911)
	20%	0.2703 (0.0859)	0.2375 (0.1266)	0.4359 (0.0743)	0.1306 (0.0812)	0.4323 (0.0589)	0.0829 (0.0857)	0.4395 (0.0515)	0.0782 (0.0964)
	30%	0.3101 (0.0836)	0.1691 (0.1088)	0.4356 (0.0589)	0.0975 (0.0535)	0.5004 (0.0532)	0.0101 (0.0246)	0.6058 (0.0241)	0.0043 (0.0169)
Sim5	5%	0.0575 (0.0214)	0.7844 (0.0748)	0.2533 (0.0596)	0.5515 (0.1117)	0.127 (0.0397)	0.5994 (0.1127)	0.2596 (0.052)	0.5138 (0.0749)
	10%	0.072 (0.0257)	0.7346 (0.0872)	0.2545 (0.0879)	0.5235 (0.1428)	0.1986 (0.0476)	0.4692 (0.1087)	0.2403 (0.0646)	0.4556 (0.0896)
	20%	0.0766 (0.0445)	0.7239 (0.1317)	0.2459 (0.0894)	0.5493 (0.1419)	0.3454 (0.0808)	0.2477 (0.1929)	0.2975 (0.065)	0.3767 (0.1543)
	30%	0.1064 (0.0455)	0.6268 (0.1366)	0.2395 (0.0739)	0.5281 (0.1067)	0.3915 (0.0912)	0.2961 (0.1328)	0.3983 (0.0705)	0.2692 (0.1217)
Sim6	5%	0.2211 (0.0515)	0.3198 (0.1062)	0.436 (0.0846)	0.1709 (0.1114)	0.2685 (0.0822)	0.2415 (0.1381)	0.4105 (0.0712)	0.1983 (0.077)
	10%	0.2573 (0.0672)	0.2515 (0.1093)	0.4155 (0.0693)	0.1921 (0.0857)	0.3305 (0.058)	0.1857 (0.0935)	0.4068 (0.1079)	0.176 (0.0958)
	20%	0.2501 (0.0588)	0.2613 (0.1072)	0.3865 (0.0655)	0.1818 (0.0869)	0.5618 (0.0497)	0.0216 (0.0377)	0.5625 (0.0433)	0.006 (0.0146)
	30%	0.2597 (0.0626)	0.2442 (0.1064)	0.4205 (0.1032)	0.1773 (0.0745)	0.4992 (0.0903)	0.0634 (0.0811)	0.4923 (0.0607)	0.1094 (0.0676)